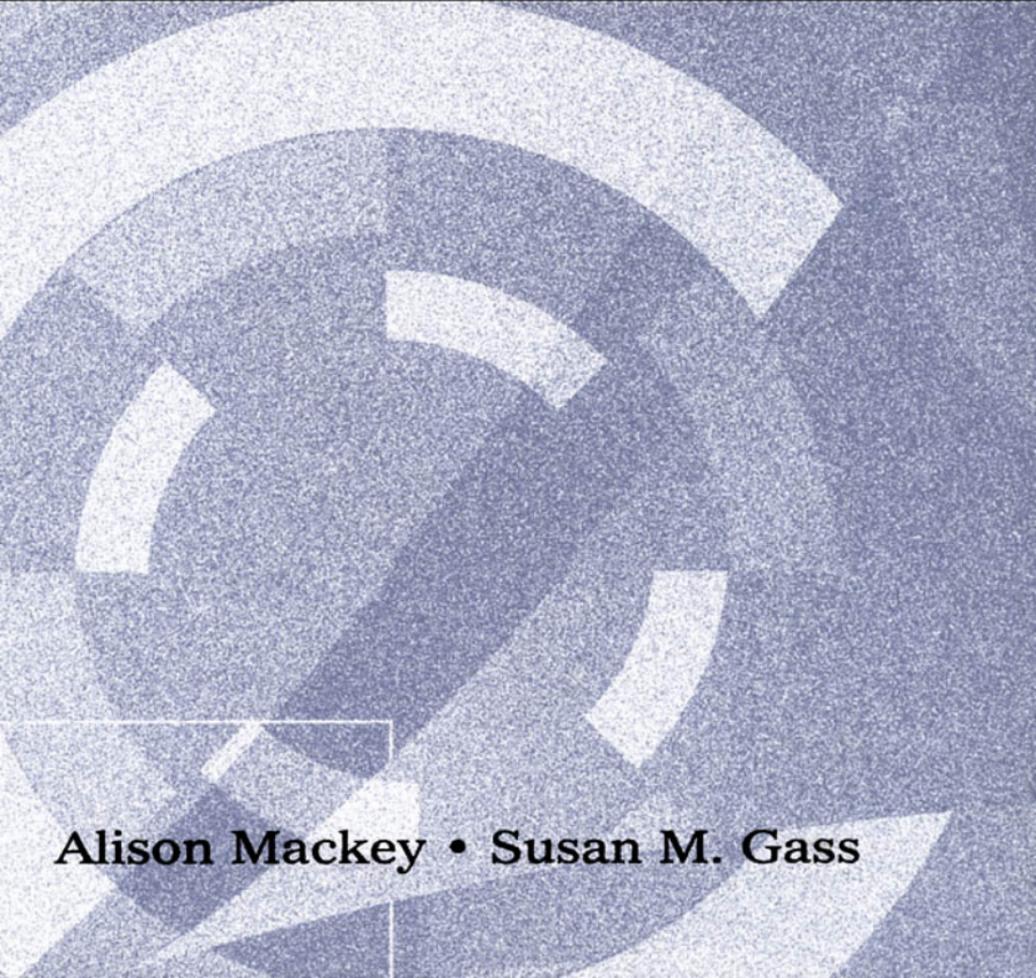


SECOND LANGUAGE RESEARCH



Methodology and Design



Alison Mackey • Susan M. Gass

**SECOND LANGUAGE
RESEARCH**

Methodology and Design

This page intentionally left blank

**SECOND LANGUAGE
RESEARCH**

Methodology and Design

Alison Mackey
Georgetown University

Susan M. Gass
Michigan State University



2005

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey

London

Copyright © 2005 by Lawrence Erlbaum Associates, Inc.

All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or any other means, without prior written permission of the publisher.

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Mackey, Alison.

Second language research : methodology and design / Alison Mackey, Susan M. Gass.

p. cm.

Includes bibliographical references and index.

ISBN 0-8058-5602-1 (cloth : alk. paper) ISBN 0-8058-4249-7 (pbk. : alk. paper)

1. Second language acquisition. 2. Second language acquisition—Research. I. Gass, Susan M. II. Title.

P118.2.M23 2005

—dc22

2004053288

CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

PREFACE	xiii
1 INTRODUCTION TO RESEARCH	1
1.1. Different Types of Research	2
1.2. What is a Research Report?	5
1.2.1. Title Page	6
1.2.2. Abstract	7
1.2.3. Introduction	7
1.2.4. Methods Section	9
1.2.5. Results	13
1.2.6. Discussion/Conclusion	15
1.2.7. Notes	15
1.2.8. References	16
1.2.9. Appendixes	16
1.3. Identifying Research Questions	16
1.3.1. Feasibility	18
1.3.2. Research Questions and Hypotheses	19
1.3.3. Replication	21
1.4. Conclusion	23
Follow-Up Questions and Activities	23
2 ISSUES RELATED TO DATA GATHERING	25
2.1. Ethical Issues In Research Involving Human Subjects	25
2.1.1. Obtaining Informed Consent From Second Language Learners	26
2.1.2. History of Institutional Review of Human Subjects Research, Compliance, and Problem Solving	36

- 2.2. Conclusion 41
- Follow-Up Questions and Activities 41

3 COMMON DATA COLLECTION MEASURES

43

- 3.1. Pilot Testing 43
- 3.2. The Significance of Data Collection Measures 44
 - 3.2.1. Syntax: Japanese Passives 45
 - 3.2.2. Interaction Research 46
 - 3.2.3. Pragmatics Research 47
- 3.3. Researching Formal Models of Language 48
 - 3.3.1. Acceptability Judgments 48
 - 3.3.2. Elicited Imitation 55
 - 3.3.3. Magnitude Estimation 56
 - 3.3.4. Truth-Value Judgments and Other Interpretation Tasks 58
 - 3.3.5. Sentence Matching 59
- 3.4. Processing Research 61
 - 3.4.1. Sentence Interpretation 61
 - 3.4.2. Reaction Time 62
 - 3.4.3. Moving Window 63
- 3.5. Interaction-Based Research 65
 - 3.5.1. Picture Description Tasks 66
 - 3.5.2. Spot the Difference 67
 - 3.5.3. Jigsaw Tasks 71
 - 3.5.4. Consensus Tasks 72
 - 3.5.5. Consciousness-Raising Tasks 74
 - 3.5.6. Computer-Mediated Research 75
- 3.6. Strategies and Cognitive Processes 75
 - 3.6.1. Observations 76
 - 3.6.2. Introspective Measures 77
- 3.7. Sociolinguistic/Pragmatics-Based Research 85
 - 3.7.1. Naturalistic Settings 86
 - 3.7.2. Elicited Narratives 87
 - 3.7.3. Discourse Completion Test (DCT) 89
 - 3.7.4. Role Play 91
 - 3.7.5. Video Playback for Interpretation 91
- 3.8. Questionnaires and Surveys 92
- 3.9. Existing Databases 97
 - 3.9.1. CHILDES 97
 - 3.9.2. Other Corpora 97
- 3.10. Conclusion 98
- Follow-Up Questions and Activities 98

4 RESEARCH VARIABLES, VALIDITY, AND RELIABILITY **100**

- 4.1. Introduction 100
 - 4.2. Hypotheses 100
 - 4.3. Variable Types 101
 - 4.3.1. Independent and Dependent Variables 103
 - 4.3.2. Moderator Variables 103
 - 4.3.3. Intervening Variables 104
 - 4.3.4. Control Variables 104
 - 4.4. Operationalization 105
 - 4.5. Measuring Variables: Scales of Measurement 105
 - 4.6. Validity 106
 - 4.6.1. Content Validity 107
 - 4.6.2. Face Validity 107
 - 4.6.3. Construct Validity 107
 - 4.6.4. Criterion-Related Validity 108
 - 4.6.5. Predictive Validity 108
 - 4.6.6. Internal Validity 109
 - 4.6.7. External Validity 119
 - 4.7. Reliability 128
 - 4.7.1. Rater Reliability 128
 - 4.7.2. Instrument Reliability 129
 - 4.8. Conclusion 130
- Follow-Up Questions and Activities 131

5 DESIGNING A QUANTITATIVE STUDY **137**

- 5.1. Introduction 137
 - 5.2. Research Materials 138
 - 5.3. Intact Classes 141
 - 5.4. Counterbalancing 143
 - 5.5. Research Design Types 145
 - 5.5.1. Correlational (Associational) Research 145
 - 5.5.2. Experimental and Quasi-Experimental Research 146
 - 5.5.3. Measuring the Effect of Treatment 148
 - 5.5.4. Repeated Measures Design 150
 - 5.5.5. Factorial Design 151
 - 5.5.6. Time-Series Design 152
 - 5.5.7. One-Shot Designs 156
 - 5.6. Finalizing Your Project 158
 - 5.7. Conclusion 159
- Follow-Up Questions and Activities 159

6 QUALITATIVE RESEARCH 162

- 6.1. Defining Qualitative Research 162
- 6.2. Gathering Qualitative Data 167
 - 6.2.1. Ethnographies 167
 - 6.2.2. Case Studies 171
 - 6.2.3. Interviews 173
 - 6.2.4. Observations 175
 - 6.2.5. Diaries/Journals 176
- 6.3. Analyzing Qualitative Data 178
 - 6.3.1. Credibility, Transferability, Confirmability and Dependability 179
 - 6.3.2. Triangulation 181
 - 6.3.3. The Role of Quantification in Qualitative Research 182
- 6.4. Conclusion 182
- Follow-Up Questions and Activities 183

7 CLASSROOM RESEARCH 185

- 7.1. Classroom Research Contexts 185
- 7.2. Common Techniques for Data Collection in Classroom Research 186
 - 7.2.1. Observations 186
- 7.3. Introspective Methods in Classroom Research 201
 - 7.3.1. Uptake Sheets 201
 - 7.3.2. Stimulated Recall 203
 - 7.3.3. Diary Research in Classroom Contexts 203
- 7.4. Practical Considerations in Classroom Research 205
 - 7.4.1. Logistical Issues to Consider When Carrying Out Classroom Research 206
 - 7.4.2. Problematics 209
- 7.5. Purposes and Types of Research Conducted in Classroom Settings 212
 - 7.5.1. The Relationship Between Instruction and Learning in Second Language Classrooms 213
 - 7.5.2. Action Research 216
- 7.6. Conclusion 219
- Follow-Up Questions and Activities 220

8 CODING 221

- 8.1. Preparing Data for Coding 221
 - 8.1.1. Transcribing Oral Data 222
- 8.2. Data Coding 225
 - 8.2.1. Coding Nominal Data 226
 - 8.2.2. Coding Ordinal Data 227
 - 8.2.3. Coding Interval Data 229
- 8.3. Coding Systems 230
 - 8.3.1. Common Coding Systems and Categories 231
 - 8.3.2. Custom-Made Coding Systems 234
 - 8.3.3. Coding Qualitative Data 241
- 8.4. Interrater Reliability 242
 - 8.4.1. Calculating Interrater Reliability 243
- 8.5. The Mechanics of Coding 246
 - 8.5.1. How Much to Code? 247
 - 8.5.2. When to Make Coding Decisions? 248
- 8.6. Conclusion 248
- Follow-Up Questions and Activities 248

9 ANALYZING QUANTITATIVE DATA 250

- 9.1. Introduction 250
- 9.2. Descriptive Statistics 250
 - 9.2.1. Measures of Frequency 251
 - 9.2.2. Measures of Central Tendency 254
 - 9.2.3. Measures of Dispersion 258
- 9.3. Normal Distribution 261
- 9.4. Standard Scores 263
- 9.5. Probability 264
- 9.6. Inferential Statistics 269
 - 9.6.1. Prerequisites 269
 - 9.6.2. Parametric Versus Nonparametric Statistics 271
 - 9.6.3. Parametric Statistics 272
 - 9.6.4. Nonparametric Tests 278
- 9.7. Statistical Tables 280
- 9.8. Strength of Association 282
- 9.9. Eta^2 and Omega^2 282
- 9.10. Effect Size 282

- 9.11. Meta-Analyses 283
- 9.12. Correlation 284
 - 9.12.1. Pearson Product-Moment Correlation 286
 - 9.12.2. Spearman Rho/Kendall Tau 290
 - 9.12.3. Factor Analysis 290
- 9.13. Statistical Packages 291
 - 9.13.1. SPSS 291
 - 9.13.2. VARBRUL 291
- 9.14. Conclusion 292
- Follow-Up Questions and Activities 292

10 CONCLUDING AND REPORTING RESEARCH 297

- 10.1. The Importance of Reporting Research 297
- 10.2. The Final Stages in Reporting Quantitative Research 298
 - 10.2.1. The Discussion 298
 - 10.2.2. Limitations, Future Research, and Conclusion Sections 302
- 10.3. The Final Stages in Reporting Qualitative Research 304
- 10.4. Reporting Combined Method (Quantitative and Qualitative) Research 307
- 10.5. Checklist for Completing Reports of Research 308
 - 10.5.1. The Research Problem and Questions 308
 - 10.5.2. The Research Hypotheses 309
 - 10.5.3. The Audience 309
 - 10.5.4. The Abstract 310
 - 10.5.5. The Literature Review 310
 - 10.5.6. The Design of the Study 311
 - 10.5.7. Logistics 311
 - 10.5.8. Participants 312
 - 10.5.9. Data Gathering 312
 - 10.5.10. Data Analysis 312
 - 10.5.11. Conclusions 314
 - 10.5.12. References 315
 - 10.5.13. Footnotes, Endnotes, Figures, and Tables 316
 - 10.5.14. Author's Note/Acknowledgments 317
 - 10.5.15. Postresearch Concerns 318
 - 10.5.16. Final Touches and Formatting 318
- 10.6. Conclusion 320
- Follow-Up Questions and Activities 320

APPENDIX A: SAMPLE SHORT FORM WRITTEN CONSENT DOCUMENT FOR SUBJECTS WHO DO NOT SPEAK ENGLISH	322
APPENDIX B: SAMPLE CONSENT FORM FOR A STUDY IN A FOREIGN LANGUAGE CONTEXT	323
APPENDIX C: SAMPLE CONSENT FORM FOR A CLASSROOM STUDY	324
APPENDIX D–G: SAMPLE INSTITUTIONAL REVIEW BOARD APPLICATION: GEORGETOWN UNIVERSITY, FORMS 1–4	326
APPENDIX H: SAMPLE TRANSCRIPTION CONVENTIONS: “JEFFERSONIAN” TRANSCRIPTION CONVENTIONS	342
APPENDIX I: SAMPLE TRANSCRIPTION CONVENTIONS FOR THE L2 CLASSROOM	345
APPENDIX J: COMMONLY-USED FORMULÆ	347
GLOSSARY	350
REFERENCES	370
AUTHOR INDEX	387
SUBJECT INDEX	393

This page intentionally left blank

Preface

This book addresses issues of research methodology. It is designed to be used as a textbook for introductory courses on research methodology and design, as well as for general courses in second language studies in which there is an emphasis on research. We have aimed to create a text that can also be used as a resource by those carrying out many different types of second language research.

We approached the book with novice researchers in mind. For this reason, we explain key concepts and provide concrete examples wherever possible for those with little or no research experience. However, we also assume that our readers will have some background in the topic of second language learning. The discussion and data-based questions and activities at the end of each chapter are aimed to promote better understanding of the concepts as readers work through the book. We also include a detailed glossary to aid researchers who prefer to use the book more as a resource than a text.

We have tried to take a broad and inclusive view of what is meant by 'second language' research. For this reason, our examples reflect concepts from a variety of perspectives in the second language research field. The book is designed to address issues important for research in both second and foreign language settings, child second language learning, bilingual language learning, as well as the acquisition of second and subsequent languages. We have attempted to cast a similarly wide net in our coverage of topics; for example, we include research design issues that range from the use of highly experimental data elicitation tools to qualitative concerns to teacher-initiated research in classrooms. We also include topics of recent interest in the field, such as dealing with university, institutional, and school review boards that grant permission for data gathering from human subjects. Although our goal is to acquaint readers with the basic issues, problems, and solutions involved in conducting second language re-

search, we believe that some of the content of the book is also relevant to a wider applied linguistics context. In other words, some issues of design are common to many areas of applied linguistics research, even though a particular example may not always be.

Although the book focuses specifically on issues of research design and methodology, we have included one chapter that focuses on statistics. Because the field of statistics is so broad and has its own specialized texts and courses, we provide only a simple overview of some of the basic concepts in this area. For those who intend to conduct detailed statistical analyses, we recommend coursework, expert consultations, and other comparable means of learning about advanced statistics, including statistics textbooks. We do not include specific recommendations about particular statistics texts because the selection of the text depends on the focus of the research problem. Second language research can focus on educational or pedagogical practice or on theory building; it can address issues from a variety of perspectives, including psychology, sociology, linguistics, and bilingualism. We suggest that users of this book consult one of the many appropriate statistics books available.

It is always difficult to decide on the order in which to present information. One researcher's ordering of material and chapters might not coincide with the preferences of another researcher or reader. We have placed information on data gathering at the beginning of the book due to the fact that our experience in teaching research methods courses over the years has led us to believe that researchers need to think about where data come from at the outset of a project, and also to think about how data are gathered before becoming immersed in some of the more technical issues of design. In this book, then, issues of data gathering serve as an anchor for later chapters. Of course, when using the book as a text, we hope that instructors will adapt the book and reorder chapters to match their particular syllabus and preference for presentation. For this reason, we have aimed for each chapter to work as a standalone introduction to the area it covers.

We are grateful to many individuals for their support in this project that ended up, like most projects of this sort, having a longer history than we had originally anticipated. We first thank the many students we have had in different classes over the years who have not hesitated to provide feedback on our various syllabi and our sequencing of materials as well as the designs of our own research. Rebekha Abbuhl and Ildikó Svetics made many valuable contributions to the process, including library work, feedback, and editing, always providing careful attention to content and detail throughout. Several reviewers also provided us with numerous useful ideas and suggestions on our proposal. We greatly appreciated the time and effort that went

into these reviewer comments. For their helpful input on this general project, Alison Mackey thanks the following students who took the research methods class at Georgetown University: Seon Jeon, Cara Morgan, and Harriet Wood. We are also particularly grateful to Rebecca Adams, Kendall King, Kimberly McDonough, Jenefer Philp, Charlene Polio, Rebecca Sachs, and Ian Thornton for help with various aspects of drafts of different chapters. Zoltán Dörnyei, Rod Ellis, and Patsy Lightbown read the entire manuscript, and their recommendations led to numerous improvements. Finally, our editor, Cathleen Petree of Lawrence Erlbaum Associates, has been unwavering in her support of this book, and we thank her.

—*Alison Mackey*
Columbia, MD

—*Susan Gass*
Williamston, MI

This page intentionally left blank

Introduction to Research

What is meant by research, and how do we identify good research questions? These are questions that are not always easy to answer, but we anticipate that by the end of this book you will be in a better position to think about them. This book is intended to be practical in nature, aimed at those who are involved in second language studies and second/foreign language teaching. We recognize that many people are often put off by the word *research*, including teachers who have been teaching for quite some time but are not involved in research, and those who are just beginning in the field. We hope to demystify the process.

The *American Heritage College Dictionary* defined *research* as “scholarly or scientific investigation or inquiry” or as a verb “to study (something) thoroughly” (2000). Thus, in its most basic and simplest form, research is a way of finding out answers to questions.

We begin by reminding the reader that we are all involved in research every day. For example, consider what is probably part of many of our lives—being stuck in a traffic jam. As we find ourselves not moving on a freeway, we ask why this has happened and come up with a hypothesis (e.g., because there is an accident ahead, or because it is 5:00 P.M. on a Friday afternoon). We then seek verification of our hypothesis by waiting patiently (or impatiently) until the traffic starts moving again. If we see an accident or the flashing lights of an emergency vehicle, we can confirm or at least strengthen our hypothesis. In the absence of an accident, we might conclude that it must be typical rush hour traffic. In other words, every day we ask questions, come up with hypotheses, and seek confirmation of those hypotheses.

In this chapter, we outline what readers can expect from a typical research report and discuss the process of generating research questions and formulating hypotheses. We conclude the chapter by discussing issues of feasibility and the importance of replication in second language research.

1.1. DIFFERENT TYPES OF RESEARCH

There are many approaches to dealing with research. Two of the most common are known as *quantitative* and *qualitative*, although this distinction is somewhat simplistic as the relationship is best thought of as a continuum of research types. Quantitative research generally starts with an experimental design in which a hypothesis is followed by the quantification of data and some sort of numerical analysis is carried out (e.g., a study comparing student test results before and after an instructional treatment). Qualitative studies, on the other hand, generally are not set up as experiments; the data cannot be easily quantified (e.g., a diary study in which a student keeps track of her attitudes during a year-long Japanese language course), and the analysis is interpretive rather than statistical. As mentioned previously, this is an overly simplistic view because one can imagine a number of variations on this theme. In general, however, quantitative and qualitative research can be characterized as shown in Table 1.1 (based on Reichardt & Cook, 1979).

In this book we attempt to be as inclusive as possible and cover as many research types as possible.

Grotjahn (1987) pointed out that there are many parameters that can be used to distinguish research types, including the type of data (quantitative or qualitative), the method of analysis (interpretative or statistical), and the

TABLE 1.1
Characteristics of Quantitative and Qualitative Research

<i>Quantitative Research</i>	<i>Qualitative Research</i>
• Obtrusive, involving controlled measurement	• Naturalistic and controlled observation
• Objective and removed from the data	• Subjective
• Verification oriented, confirmatory	• Discovery oriented
• Outcome-oriented	• Process oriented
• Reliable, involving "hard" and replicable data	• "Soft" data
• Generalizable	• Ungeneralizable, single case studies
• Assuming a stable reality	• Assuming a dynamic reality
	• Close to the data

manner of data collection (experimental or nonexperimental [naturalistic]). He outlined six “mixed” forms, as shown in Table 1.2.

To understand the two ends of the continuum—namely “purely” quantitative and “purely” qualitative studies—consider the following abstracts of two research reports.

Quantitative Research

Interaction has been argued to promote noticing of L2 form in a context crucial to learning—when there is a mismatch between the input and the learner’s interlanguage (IL) grammar (Gass & Varonis, 1994; Long, 1996; Pica, 1994). This paper investigates the extent to which learners may notice native speakers’ reformulations of their IL grammar in the context of dyadic interaction. Thirty-three adult ESL learners worked on oral communication tasks in NS–NNS pairs. During each of the five sessions of dyadic task-based interaction, learners received recasts of their nontargetlike question forms. Accurate immediate recall of recasts was taken as evidence of noticing of recasts by learners. Results indicate that learners noticed over 60–70% of recasts. However, accurate recall was constrained by the level of the learner and by the length and number of changes in the recast. The effect of these variables on noticing is discussed in terms of processing biases. It is suggested that attentional resources and processing biases of the learner may modulate the extent to which learners “notice the gap” between their nontargetlike utterances and recasts. (Philp, 2003, p. 99)

This description meets the criteria of a quantitative study: it has quantitative data, it analyzes the data and provides results based on statistics, and the data were collected experimentally.

Qualitative Research

This ethnographic report “thickly describes” (Geertz, 1973) the participation of ESL children in the daily classroom events of a mainstream first-grade classroom. Data for this paper come from a year-long study of one classroom in an international school on a college campus in the U.S. Using a language socialization and micropolitical orientation, the report describes how, through socially significant interactional routines, the children and other members of the classroom jointly constructed the ESL children’s identities, social relations, and ideologies as well as their communicative competence in that setting. The sociocultural ecology of the community, school, and classroom shaped the kinds of microinteractions that occurred and thus the nature of their language learning over the course of the year. (Willett, 1995, p. 473)

TABLE 1.2
Six Mixed Forms of Research

<i>Type of Research</i>	<i>Form of Data</i>		<i>Method of Analysis</i>		<i>Manner of Data Collection</i>	
	<i>Quantitative</i>	<i>Qualitative</i>	<i>Statistical</i>	<i>Interpretative</i>	<i>Experimental/ Quasi-Experimental</i>	<i>Nonexperimental</i>
Experimental-qualitative-interpretative		✓		✓	✓	
Experimental-qualitative-statistical		✓	✓		✓	
Experimental-quantitative-interpretative	✓			✓	✓	
Exploratory-qualitative-statistical		✓	✓			✓
Exploratory-quantitative-statistical	✓		✓			✓
Exploratory-quantitative-interpretative	✓			✓		✓

This abstract uses naturalistic data (observations of students in a classroom), provides an interpretative rather than a statistical analysis, and uses a nonexperimental design. We address a spectrum of issues related to qualitative research in chapter 6.

1.2. WHAT IS A RESEARCH REPORT?

In this section, we provide a guide for readers as to what to expect in a typical article in the second language research field, focusing primarily on quantitatively oriented research articles. Unlike quantitative research reports, for which there is a relatively standard format for reporting, qualitative research articles are more wide ranging in terms of organization (for more information, see chapter 6, in which we discuss qualitative research).

In this chapter our goal is to give an idea of what to expect in a research report. To that end, following is a basic skeleton of a research paper. (Chapter 10 provides detailed information for researchers concerning the *writing* and *reporting* of their own research based on all of the areas covered in this book.)

Typical Research Paper Format

TITLE PAGE

ABSTRACT

BODY

- I. Introduction
 - A. Statement of topic area
 - B. Statement of general issues
 - C. General goal of paper
 - D. Literature review
 1. Historical overview
 2. Major contributions to this research area
 3. Statement of purpose, including identification of gaps
 4. Hypotheses
- II. Method
 - A. Participants
 1. How many?
 2. Characteristics (male/female, proficiency level, native language, etc.)
 - B. Materials

1. What instruments?
 2. What sort of test? What sort of task?
- C. Procedures
1. How is the treatment to be administered?
 2. How/when is the testing to be conducted?
- D. Analysis
- How will the results be analyzed?
- III. Results
- Charts, tables, and/or figures accompanied by verbal descriptions
- IV. Discussion/conclusion (often two separate sections)
- Common features:
- Restatement of the main idea of the study
 - Summary of the findings
 - Interpretation of the findings in light of the research questions
 - Proposed explanation of the findings, usually including information about any findings that were contrary to expectations
 - Limitations of the study
 - Suggestions for future research

NOTES

REFERENCES

APPENDIXES

We now consider in more detail what might be included in some of these parts of a typical research paper.

1.2.1. Title Page

The title page includes these elements:

- Name of author(s)¹

¹When multiple authors are involved, it is advisable to make decisions as early as possible in the research process as to whose names will be on the final version of the research report and in what order the names will appear. As the process evolves, changes might be necessary; however, to avoid difficulties in the long run, it is best to make sure that there is agreement as to authorship and expectations of work wherever possible. The *Publication Manual of the American Psychological Association, Fifth Edition* put it this way: "To prevent misunderstanding and to preserve professional reputations and relationships, it is best to establish as early as possible in a research project who will be listed as an author, what the order of authorship will be, and who will receive an alternative form of recognition" (American Psychological Association, 2001, pp. 6–7).

- Title of paper
- Contact information

1.2.2. Abstract

The abstract presents a summary of the topic of the paper and the major findings of the research. Abstracts are very often printed through abstracting services and are generally the primary step in finding out about a paper. They are usually 100–150 words in length, although there is variation depending on where the article is published. Following is an example of an abstract:

Abstract

Recent studies have suggested that the incorporation of some attention to form into meaning-centered instruction can lead to improved performance in processing input and increased accuracy in production. Most have examined attention to form delivered by instructors or instructional materials. This study examines the production of 8 classroom learners at 4 levels of proficiency to determine the extent to which learners can and do spontaneously attend to form in their interaction with other learners. Results suggest that the degree and type of learner-generated attention to form is related to proficiency level and the nature of the activity in which the learners are engaged. They also indicate that learners overwhelmingly choose to focus on lexical rather than grammatical issues.

(118 words; from Williams, 1999, p. 583)

In this short abstract, two sentences are devoted to past research, with the third sentence informing the reader what this study is about and how it fills a gap in the literature. The final two sentences provide information about what the reader can expect from the results.

1.2.3. Introduction

The introduction sets the scene and provides the reader with background material (statement of topic area and general issues) as well as an outline of the purpose of the research. This is generally followed by a literature review. Some possibilities for literature reviews include the following:

- Historical overview.

Example: In earlier views of the relationship between x and y ...

- Major players in this research area, including questions, past findings, and controversies.

Example: In 1998, Ellis claimed that the relationship between x and y was an important one and went on to show that ...

However, in a more recent paper, Zhang (1995) argued that this relationship could not be valid because ...

- General goal of the paper.

Example: In this paper I will argue that Zhang's interpretation of Ellis's data is incorrect and that when one looks at variable z in the context of x and y , the relationship is indeed valid. I will present data that support Ellis's original interpretation of abc .

- Research questions/hypotheses. In Williams' (1999) article abstracted earlier, the following research questions are provided after the introduction (p. 591):

Example:

1. Do learners in learner-centered, communicative classrooms spontaneously attend to form?
2. Is proficiency level related to the extent to which they do so?
3. How do learners draw attention to form?
4. When do learners draw attention to form, that is, during what types of activities?
5. What kinds of forms do they attend to?

As can be seen, these questions build on one another. They are not, however, formulated as predictions. Following are some of the specific hypotheses from a different study (Gass & Alvarez-Torres, 2005):

Example:

1. Given that interaction is said to be an attention-drawing device, we predict that the three experimental groups with interaction will perform better than the group with no interaction.
2. Because input and interaction serve different important functions, when there is a combination of conditions (input followed by interaction and interaction followed by input), performance will be better than when only one type of presentation is available.

3. Given Gass's (1997) assumption that interaction serves as a priming device that "readies" learners to utilize follow-up input, the best performance will take place in the group with interaction followed by input.

The amount of detail needed in a literature review will depend on the purpose of the report. For example, a doctoral dissertation will generally be as exhaustive as possible. On the other hand, the literature review for a journal article or for a chapter in a book will only address previous research that directly relates to the specific purpose of the research report and might only be about 8–10 pages.

1.2.4. Methods Section

In the methods section, the reader can expect to be informed about all aspects of the study. One reason for this is the later possibility of replication (see section 1.3.3.). Another reason is that in order for readers to come to an informed opinion about the research, they need to know as much detail as possible about what was done.

1.2.4.1. Participants

This section includes information about the participants² in a study. For example, how many participants were there? What are their characteristics (e.g., male/female, native language, age, proficiency level, length of residence, amount and type of instruction, handedness)? The characteristics that researchers describe will depend, in part, on the experiment itself. For example, handedness was listed as a possible characteristic. This would probably be relevant in a study that required participants to press a button on a computer as a response to some stimulus. Most such studies are set up for right-handed individuals, so it might be important to know if the particular setup favored those individuals.

1.2.4.2. Materials

The materials used to conduct the study are usually presented in detail. Below is an example of a materials section from an article on deriving

²According to the *Publication Manual of the American Psychological Association, Fifth Edition*, the word *participant* is more appropriate than words such as *subject*. In section 2.12, they stated, "Replace the impersonal term *subjects* with a more descriptive term when possible and appropriate—*participants, individuals, college students, children, or respondents*, for example" (American Psychological Association, 2001, p. 65). They went on to say that when discussing statistical results, the word *subjects* is appropriate.

meaning from written context by Dutch children (grades 2, 4, and 6) in their L1 (Fukkink, Blok, & de Gloppe, 2001):

Target words were selected from a primary-school dictionary (Verburg & Huijgen, 1994), to warrant that relevant concepts would be selected, representative of the words young readers encounter during reading. An initial sample of words with a frequency below 10 per million (Celex, Centre for Lexical Information, 1990) was selected from this dictionary to ensure that no words were used that students were already familiar with. Three judges evaluated the concreteness of the target words, defined as a dichotomy, and words were excluded if the judges did not arrive at a unanimous agreement. A final sample of 12 words was selected, evenly divided into concrete and abstract words. The average word frequency of the words in the sample is 4.4 per million (ranging from 1 to 10 per million). Only morphologically nontransparent words were included, to promote deriving word meaning from (external) context.

Short texts of approximately a hundred words were constructed for each target word. The difficulty level of each text was adjusted to an appropriate level for average readers at the end of grade 2 on the basis of a reading difficulty index (Staphorsius & Verhelst, 1997). The narrative texts contained no explicit clues (e.g., synonyms, antonyms, or description clues). Target words were not placed in the first sentences of the text.

A version of the twelve texts was presented to three adults with target words deleted. They were instructed to fill each cloze with an answer that was as specific as possible and fitted the context. Only four out of the 36 answers, each concerning a different target word, did not match the concept of the deleted word. The other answers, however, were identical or synonymous with the deleted target word (58%) or closely related hypernyms (31%) ("to break" was filled in for the deleted target word "to shatter," for example). The texts were therefore considered to provide sufficient contextual support. (Fukkink et al., 2001, p. 481)

As can be seen, there is sufficient information provided for the reader to understand the nature of the task that these learners were being asked to do.

In addition to treatment materials, assessment materials may also appear in this section or, alternatively, this section may be divided into two sections, one dealing with treatment materials and another with testing/assessment materials. An example of assessment materials from a study on think alouds and reactivity follows. The authors measured comprehension, intake, and written production following a think-aloud task. Only a portion

of the description for each measure is provided. In all three instances, the actual tool is provided in the Appendix of their article.

To measure participants' comprehension, an 11-item comprehension task was designed to elicit 17 pieces of information based exclusively on the advice, tips, or recommendations provided through the imperatives found in the text. The information was elicited predominantly via short and multiple-choice answers

To measure participants' intake of the targeted forms, a multiple-choice recognition task was prepared. The 17 items on this task were also based exclusively on the advice, tips, or recommendations provided through the imperatives found in the text

To measure participants' controlled written production of the targeted forms, a fill-in-the-blank task, comprising 17 items that provided a list of advice for leading a healthy life, was prepared (Leow & Morgan-Short, 2004, p. 45)

The materials section presents a description of the actual materials used, but does not specify how they were used. The procedures section provides that information.

1.2.4.3. Procedures

The next questions that a reader can expect to be informed of include logistical issues related to what was actually done. How exactly was the task carried out? How was the treatment administered? How and when was testing done? Following is the procedures section from the Fukkink et al. (2001) study discussed previously:

Participants were tested individually. Sessions started with a standardized explanation of directions to the students. It was decided that each text would first be read orally by the student, because reading aloud first appeared to encourage giving oral definitions in a pilot study and a previous study (Van Daalen-Kapteijns, Elshout-Mohr, & de Gloppe, 2001). Students tried to decipher the meaning of the target word thereafter in response to the question, "Which explanation does the dictionary give for this word?" Students were permitted to reread the text.

A warming-up task was introduced first, using materials that were similar to the experimental task. The experimental items were introduced only if students demonstrated adequate understanding of the proce-

ture. The order of items was randomized for each participant. The sessions were tape recorded and transcribed for coding. (Fukkink et al., 2001, p. 482)

The Fukkink et al. (2001) study contained a separate section for scoring, in which detail was provided as to how responses were scored. A subsequent analysis section presented information about the statistical procedures used to analyze the data.

1.2.4.4. Analysis

In some research reports, the mode of analysis may be a separate section or may be included in the results section. We present two examples of what might be included in a discussion of how one will analyze the results. The first, from Leow and Morgan-Short (2004), provides information about the scoring procedure. The second, a study on planning and narrative writing from Ellis and Yuan (2004) presents information about the statistical procedures to be used.

Scoring Procedure

For the recognition and controlled written production tasks, one point was awarded to each correct answer, and no points for incorrect answers, for a total of 17 points. The comprehension task was scored in the following manner: For all items except item 1, one point was awarded for each correct answer and zero for an incorrect one. For item 1, five out of seven correct responses were required before one point was awarded. For item 11, answers could have been provided in either English or Spanish. (Leow & Morgan-Short, 2004, p. 46)

Data Analysis

The normal distribution of the three groups' scores on all variables was tested in terms of skewness and kurtosis. A series of one-way ANOVAs were subsequently performed followed by post hoc Scheffé tests where appropriate (i.e., if the F score was statistically significant). In the one variable where normal distribution was not evident ..., a Kruksal-Wallis Test was run, followed by independent *t*-tests to compare the pairs of groups. The alpha for achieving statistical significance was set at .05. Additionally, effect sizes were calculated ... to examine the size of the effect of the different kinds of planning on performance of the task (Ellis & Yuan, 2004, p. 72)

It is not always the case that all of these categories appear in every research report. Some may be combined, and others may not be relevant.

The precise organization of the report will depend on the design of the study and the authors' preference for presentation of the data.

1.2.5. Results

In this section of a research article, the results are presented with verbal descriptions of data that are also often displayed in charts, figures, or tables. Results sections usually provide objective descriptions presented without interpretation. The excerpt that follows is a small part of a results section from Philp (2003).

The provision of recasts depended entirely on the production of non-targetlike forms by each learner. Generally, as illustrated in Table 2, each learner received 44–55 recasts of question forms over five sessions with those in the Low group generally receiving higher numbers of recasts. Of these recasts, all groups received over 60% of recasts of stage 4 questions.

As shown in Table 3, the High group was presented more frequently with long recasts (62%), whereas the Low group received more short recasts

TABLE 2
Recasts Provided to Each Group

<i>Group</i>	<i>n</i>	<i>Recasts</i>		<i>Percentage of Question Forms in Recasts</i>		
		<i>N</i>	<i>M</i>	<i>Stage 3</i>	<i>Stage 4</i>	<i>Stage 5</i>
High	15	659	43.93	7 (44)	65 (415)	28 (179)
Intermediate	11	531	48.93	8 (42)	62 (316)	30 (155)
Low	7	379	54.14	6 (15)	63 (237)	33 (122)

TABLE 3
Length of Recasts and Changes to Learners' Utterances in Recasts: Percentages by Group

<i>Group</i>	<i>n</i>	<i>Length of Recast</i>		<i>Number of Changes</i>		
		<i>Short</i>	<i>Long</i>	<i>1 Change</i>	<i>2 Change</i>	<i>≥ 3 Changes</i>
High	15	38	62	39	30	31
Intermediate	11	52	48	37	31	32
Low	7	67	33	30	30	40

(67%). Similar numbers of recasts with one, two, or three or more changes to the learner's trigger utterance were received by all groups, although the Low group received slightly more of the latter. A comparison between groups is shown in Fig. 1.

Additional information about statistical results is also presented in the results section, as seen here:

Results

To test hypothesis 1, which predicted that recall of recasts would be more accurate the higher the level of the learner, the High, Intermediate, and Low groups were compared. The results of a one-way ANOVA ... show a significant effect for learner level on recall of recasts. With an alpha level of .05, the effect of learner level on recall of recasts was statistically significant, $F(2,30) = 4.1695$, $p < .05$. A priori contrasts, tested by the *t*-statistic, were computed to establish the source of difference between groups. A significant difference was found between the High and Intermediate groups on the one hand and the Low group on the other ($p < .05$). The High and Intermediate groups were not significantly different in performance on recall ($p = .814$) (Philp, 2003, p. 111)

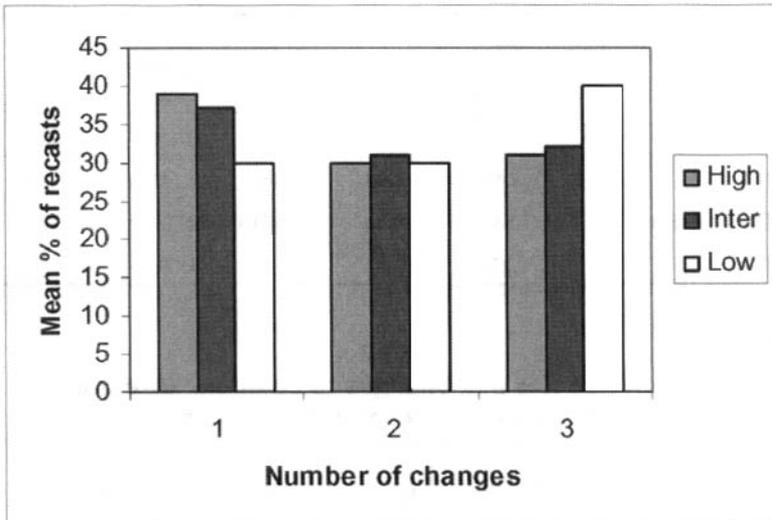


FIG. 1. Comparison by group of proportion of number of changes in recasts. From Philp, J. (2003). Constraints on 'noticing the gap': Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition*, 25, 110. Copyright © 2003 by Cambridge University Press. Reprinted with the permission of Cambridge University Press and with the permission of J. Philp.

1.2.6. Discussion/Conclusion

The discussion and conclusion are often two separate sections and are primarily interpretive and explanatory in nature. The main idea of the study may be restated and the findings summarized. Then, the findings are interpreted in light of the research questions and an explanation is attempted (perhaps with regard to findings that were contrary to expectations). Following is an example from a discussion section on form-meaning mapping by native and nonnative speakers (Jiang, 2002) in which the author, in three separate paragraphs, provided a summary, an interpretation (along with problems), and a possible explanation. We reproduced the first sentences from each paragraph:

Summary Statement

The results of experiment 1 show that whether an L2 word pair shares a single L1 translation does not affect native speakers' performance in the rating task

Interpretation and Problems

Although the findings of experiment 1 are consistent with the L1 lemma mediation hypothesis, there are two potential problems that have to be resolved before one can interpret the findings as evidence in support of it

Explanation

One explanation for this discrepancy may lie in the possible involvement of conscious knowledge about L2 words in the rating task on the part of the nonnative speakers (Jiang, 2002, p. 624)

Finally, many studies include a section on the limitations of the study and suggest ways of remediating the limitations. Possible topics for future research may also be included. Typical contents of discussion, conclusion, and limitations sections are also discussed at length in chapter 10, in which we provide tips on writing and reporting research.

1.2.7. Notes

In some journals, any parenthetical material in an article is placed in footnotes at the bottom of the relevant page. In other journals, this material may appear as endnotes—a section in which all the notes are col-

lected together at the end of the article. In addition, there is generally an Author's Note, often including contact information, information concerning prior presentations based on the research presented in the paper, and acknowledgments.

1.2.8. References

In most journals in the second language research field, everything cited in the paper appears in the reference list, and all sources listed in the reference list are cited in the paper. There is no single style used by all journals in the field; different journals have different styles for references. The use of style manuals is further discussed in chapter 10.

1.2.9. Appendixes

The appendixes to a research article may include examples of the actual materials used in the study, along with any other information that, although necessary for the interpretation of the study, might interrupt the flow of the paper if included in the body of the article.

In this section, we have provided a brief description of what can be expected in a typical quantitatively oriented article in the field of second language research. We now move on to the main focus of this book, which is how to do second language research. We begin by considering the identification of research questions.

1.3. IDENTIFYING RESEARCH QUESTIONS

The first question and perhaps one of the most difficult aspects of any research undertaking is the identification of appropriate research questions. Research questions are an integral part of quantitative research. The identification process for qualitative research, discussed in chapter 6, is often quite different than for quantitative research. For example, in qualitative studies, in keeping with the goals of research, questions are often not as narrowly constrained as they are in quantitative studies.

Questions need to be interesting in the sense that they address current issues; at the same time, they need to be sufficiently narrow and constrained so that they can be answered. Broad questions can be difficult if not impossible to address without breaking them down into smaller answerable questions. For example, a general research question such as "What is the effect of the native language on the learning of a second or foreign language?" cannot be an-

swered as formulated. This is because it represents a research area, but not a specific research question. To address the research area, a researcher might investigate the effect of a native language on specific aspects of a target language (e.g., phonology, syntax). One way to begin to reduce the general question would be to consider the learning of a language that has a linguistic category not present in the native language. Again, this is somewhat broad, so the researcher might want to further reduce this to a specific question: "How do learners of a nontonal language learn to make lexical distinctions using tone?" This is a reasonable starting point for the investigation of this question. The researcher could then examine the interlanguages of native speakers of English learning Chinese. Of course, the researcher would have to determine whether he or she wanted to examine production or comprehension in order to come up with specific hypotheses and a design. We return to the issue of hypotheses in section 1.3.2 of this chapter.

From where do research ideas come? We mentioned earlier that research investigations need to be current, which of course entails that the questions have not already been answered in the literature, or have only partially been answered and therefore require further or additional investigation. Research questions also need to be theoretically interesting; otherwise, we run into a "so what" response to our research. Most reasoned research questions come from a reading of the literature and an understanding of the history of current issues. The conclusion sections of many articles suggest questions for future research. Some are quite specific, whereas others are merely suggestive. Following are some examples from journals. The first is a study of lexical repetition as a function of topic, cultural background, and development of writing ability by learners of English who are native speakers of Arabic, Japanese, Korean, and Spanish. The second comes from a study on the acquisition of English causatives by native speakers of Hindi-Urdu and Vietnamese. The third is a study of the relationship between speech and reading in a group of ESL learners who are native speakers of Japanese, Chinese, Korean, and Farsi.

Examples From Studies That Suggest Questions for Future Research

Future studies may wish to examine other possible topic-related variations, including distinctions between personal and nonpersonal writing and among different writing purposes. A second question is whether the time limitation imposed on these essays encouraged the use of repetition as a cohesion strategy. (Reynolds, 2001, p. 472)

There is, nonetheless, a need for further research in this area, involving a larger repertoire of verb classes, as well as a wider range of proficiency levels.... Similarly, further research could be undertaken on the

influence of L1 verb serialization in languages like Vietnamese on the acquisition of the argument structure of verbs in nonserializing languages like English.... Further research could also include studies on the acquisition of semantic classes relevant to various syntactic phenomena, involving a variety of languages (both as L1s and L2s), with different morphologies, classes of verbs and selectional restrictions on verbs. (Helms-Park, 2001, p. 94)

... [F]urther exploration of the alphabet group differences on the figure pairs might prove productive. This should include a more detailed analysis of three groups (Roman alphabet, non-Roman alphabet, and Ideographic) instead of two groups as in the present study. In addition, the students might be presented with the decision tasks in their native language as a further control against the test effects (Muchisky, 1983, pp. 94–95)

As mentioned earlier, we often develop research questions through suggestions made by other researchers. Another way is through the extensive reading and analysis of existing research, which can lead to the identification of gaps that may strike a reader as important. Often, when reading an article, one might recognize that something has not been controlled for or that different languages might function differently on a certain important dimension. Alternatively, some controversy may have been left unresolved. This information may turn out to form the basis of a follow-up study, but a researcher must first make sure that others have not conducted such studies. A first step in this process is to consult a citation index (see your university librarian) to locate work that has cited the paper on which you will be basing your study. Another way of locating relevant information is through Web-based searches, which often yield studies published in a range of venues.

On other occasions, ideas for research might stem from observing learners either in or out of a classroom context or through some general feeling of curiosity having observed nonnative speaker linguistic behavior. These ideas may or may not develop into research studies, but, in any case, the first task is to conduct a literature search to see what has already been done. There are many databases available for this purpose. Again, university librarians can assist with this process, and Web-based searches can often yield fruitful results.

1.3.1. Feasibility

The feasibility of a study may depend on a number of factors, some of which we have already mentioned (e.g., the breadth of the study in relation

to its research questions' scope and answerability). Another factor to take into account when considering feasibility is whether or not it will be possible to obtain the data necessary to answer the question. Consider a study in which one wants to conduct a survey of the attitudes of heritage learners (i.e., students who are learning the language of their parents, grandparents, etc.). In order to do this, the researcher first has to define exactly what constitutes a heritage learner. One question might be whether someone can be considered a heritage learner if she or he has distant relatives in Uzbekistan, for example, but has only very rarely heard the language spoken. Following this step, the researcher needs to go about identifying individuals who would qualify under the definition chosen. In many settings, it would be difficult to find a reasonable number of participants to make the study interesting. Thus, pertinent data sources need to be identified as a part of determining the feasibility of the study.

Another study might seek to compare performance on different communication task types. As we discuss in chapter 3, there are many important dimensions on which communicative tasks can differ. However, it might not be feasible to require participants to do 15 different tasks. Exhaustion and boredom might set in, and the researcher would not know how to interpret the results. This is not to say that such a study could not be conducted; it is just that the design of the study might entail large numbers of participants who may or may not be available for the many rounds of data collection that such a study would necessitate.

Thus, any study should be designed with a full understanding of the fact that the limitations of the setting and the population might constrain the research.

1.3.2. Research Questions and Hypotheses

Research problems are generally expressed in terms of research questions and/or hypotheses. Research questions are the questions for which answers are being sought, whereas research hypotheses can be used to express what the researcher expects the results of the investigation to be. The hypotheses are based on observations or on what the literature suggests the answers might be. There may be times when, because of a lack of relevant literature, hypotheses cannot be generated because the researcher is dealing with new and/or unexplored areas.

The literature review that leads up to the hypotheses should report all sides of an issue. In other words, fair and complete reporting is essential in any research study. We return to the issue of hypotheses in chapter 4.

To see examples of both research questions and hypotheses, consider the following from Lakshmanan (1989). This was a study that investigated the acquisition of verb inflection and the use of pronouns by children learning English (native speakers of Spanish, French, and Japanese). The data, collected by other researchers, are from longitudinal studies of four children. Below are five research questions from this study (pp. 84–85).

- RQ#1. Do null subjects in the interlanguage (IL) of these child L2 learners decrease with time?*
- RQ#2. Is there a developmental relation between null subjects and verb inflections in the IL of these child L2 learners? In other words, is increase in verb inflections accompanied by a corresponding decrease in the use of null subjects?
- RQ#3. Are obligatory verb inflections acquired at the same time for all the categories of verb morphology or does the acquisition of verb inflections depend on the specific category of verb morphology (e.g., *be* copula, auxiliaries *be*, *do*, *have*, present 3rd singular regular, past regular etc.)?
- RQ#4. Is there a developmental relation between null subjects in *is* constructions (*is* copula and auxiliary utterances) and *is* constructions? In other words, does the proportion of null subjects present in *is* contexts increase with the increase in the proportion of *is* constructions?
- RQ#5. Are there any differences between the distribution of null subjects and subjects in *is* constructions and *non-is* constructions in these child L2 learners' IL? (Lakshmanan, 1989, pp. 84–85)

As can be seen, these research questions are expressed as explorations of relationships. Lakshmanan also formulated them as hypotheses. Examples of hypotheses stemming from these research questions are given next:

- Hypothesis 1. Null subjects in the four subjects' IL will decrease with time.
- Hypothesis 2. There is a negative relation between the development of verb inflections and the use of null subjects; in other words, null subjects will decrease with the increase in verb inflections.
- Hypothesis 3. The acquisition of obligatory verb inflections depends on the specific category of verb morphology.

*Null subjects refer to expressions in languages such as Italian or Spanish that have verbs with no overt subjects. In Italian, for example, to say *I speak Italian*, one can say *Parlo italiano*, where the first word means *I speak*. The overt word for *I (io)* is not used.

- Hypothesis 4. There is a positive relationship between the use of null subjects in *is* constructions and the development of *is* constructions.
- Hypothesis 5. There are significant differences between the distribution of null subjects and lexically realized subjects in *is* constructions and *non-is* constructions. The frequency of occurrence of null subjects will be greater than the frequency of occurrence of lexically realized subjects in *is* contexts; the frequency of occurrence of null subjects in *non-is* contexts will be less than the frequency of occurrence of lexically realized subjects in *non-is* contexts. (Lakshmanan, 1989, pp. 85–86)

1.3.3. Replication

Replication is a central part of the development of any field of inquiry. If one cannot repeat the results of a particular study, the validity of the results of the original study might be called into question.³ In fact, the *Publication Manual of the American Psychological Association, Fifth Edition* stated “The essence of the scientific method involves observations that can be repeated and verified by others” (American Psychological Association, 2001, p. 348). Likewise, Albert Valdman, the editor of the journal *Studies in Second Language Acquisition*, asserted that “the way to more valid and reliable SLA research is through replication” (1993, p. 505). As Porte (2002) further noted, without these critical replication studies, “it will be extremely difficult ever to discover the definitive response to a research question or hypothesis found in one particular study ... which then permits us to generalize those findings to fit exactly another context of language learning” (p. 35). It is thus crucial that researchers report in enough detail to allow others to determine with precision what has been done. The journal *Language Learning* makes this explicit in their Instructions for Contributors by saying “Methods sections must be detailed enough to allow the replication of research.” Unfortunately, because much research in the field of second language learning is published in journals, space constraints often preclude full and complete reporting. To this end, Polio and Gass (1997) recommended that researchers submit detailed appendixes for publishers to keep either online

³Along with the issue of replication is the important issue of data reporting. How much should be reported? How much detail? The simple answer is: enough so that someone can replicate the study. Replication is discussed at greater length in chapters 6 and 8.

or as hard copies if journal space is limited, although publishers have not yet embraced this idea. More specifically, Polio and Gass suggested that these appendixes include information about any guidelines used for coding the data, measures of proficiency or development, instruments for data elicitation (including pre- and posttests), experimental protocols, and biodata on the participants. Generally speaking, there are two primary reasons for replication: verification and generalizability. As we point out later, in second language studies these issues are often fused.

Replication studies do not often find their way into the published pages of SLA literature. In fact, the only journal in the field of second language research that explicitly solicits replication studies is *Studies in Second Language Acquisition*. One reason behind the dearth of replication studies, as Valdman (1993) acknowledged, is that “to be sure, in replication one loses the aura of glamour and the exhilaration of innovation” (p. 505). This was echoed by van der Veer, van Ijzendoorn, and Valsiner (1994): “As these replication studies do not yield novelty, but rather check the reliability of the original results, they are less valued in a community where (limited) originality is highly valued” (p. 6). This speaks to the so-called political and career reasons for which an individual might decide not to pursue replication studies.

There are also academic reasons having to do with the difficulty involved in replication. A researcher can easily replicate the instruments, the task, and the setting of a study. But when dealing with linguistic behavior, individual characteristics such as prior linguistic background and knowledge come into play that would clearly be impossible to replicate for a variety of reasons. Polio and Gass (1997) discussed a continuum of replication studies (see also Hendrick, 1990; van der Veer et al., 1994), ranging from virtual to conceptual. Virtual replications in which everything is copied are clearly almost impossible. No group of participants is going to be “identical” to another group. However, conceptual replications are relatively realistic and can provide important supporting or disconfirming information. The researcher needs to be “conceptually” true to the original study, carefully considering the theoretical claims of the original research. If the claims cannot be extended to a different setting and to a different group of participants, then the generalizability (and, by implication, the validity) of the original study can be called into question. It is in this sense that it is difficult to interpret the results of a replication study. True replication, although possible in some sciences, is not possible in second language studies. Thus, when results diverge from the original study, we are left with two questions: Are the results different because they are not generalizable, or are they different because there is an issue of verification of the original results? This complex

issue is discussed in greater detail in chapter 4, when we turn to questions of internal and external validity and the interpretation of results.

1.4. CONCLUSION

In this chapter we have dealt with some of the basics of L2 research, including the range of different types of research that exist, what to expect from a typical research report, and how to identify research questions, generate hypotheses, and consider issues such as feasibility and the role of replication in second language research. In chapter 2 we deal with the question of research ethics, focusing on the important issue of informed consent.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. In this chapter we mentioned and cited articles from a small number of journals focusing on second language research. Clearly, there are many more. Conduct a library or online search and come up with a list of 10 journals focusing on some area (general or specific) of second language research.
2. Consider the journals you listed for question 1. Can you determine the scope of each journal? With what kinds of topics do they deal? Some journals are quite explicit; others might require a look through the table of contents and abstracts.
3. Select five of these journals and consider the extent to which the articles follow the framework set up in this chapter. If they do not, in what way(s) do they deviate?
4. Consider these same five journals. Do the journals give guidelines for submission (e.g., length, style guidelines, number of copies to submit, mode of submission)? List the guidelines you have found.
5. Find abstracts from three different articles in three different journals. Analyze each in the way that we did at the end of section 1.2.2.
6. Find three articles and consider the end of the discussion section or perhaps the beginning of the conclusion section to determine if the authors acknowledged limitations of the study. What did they say?
7. Read the conclusion sections from three different articles in three different journals. Did the authors point to future research possibilities? If not, did they do this elsewhere (perhaps shortly before the conclusion)? What did they say, and are there any issues that are of interest to you?

8. How can the following research topics be turned into researchable questions?

Example:

- Gender differences in language classes
- Do males perform differently than females on a grammar test following treatment in which negative feedback is given?
 - a. Motivation
 - b. Task effectiveness
 - c. Novice teacher performance
 - d. Attention
 - e. Final grades

Issues Related to Data Gathering

In this chapter, we introduce an increasingly important issue related to gathering data from second language learners. We focus our discussion on ethical issues in research involving humans, including the process of obtaining informed consent and the institutional review of research, together with the steps to be taken in preparing a research protocol.

2.1. ETHICAL ISSUES IN RESEARCH INVOLVING HUMAN SUBJECTS¹

Second language researchers often have questions about why approval from institutions and informed consent from individuals is necessary to collect data from human subjects, given that second language research usually poses minimal to no risks and often provides added benefits, such as language production practice. To address these questions, in the next few sections we provide a brief review of the development of guidelines in the United States.²

¹As noted in chapter 1, wherever possible throughout this book we refer to *learners* or *participants* in research. However, the nature of the current chapter requires that we use the term *subjects* in the sections on research involving humans. The word *subjects* is discouraged by many style guides.

²Guidelines for research involving human subjects are available in a number of countries, although formal regulations seem to be most specific in the United States and Canada. This chapter is based mostly on U.S. sources. In Canada, there is a tricouncil policy statement on the ethical conduct for research involving humans developed by the Canadian Medical Research Council, Natural Sciences and Engineering Research Council, and Social Sciences and Humanities Research Councils. Individuals and institutions are required to comply with the policy in order to qualify for funding. The policy is fully described at: <http://www.pre.ethics.gc.ca/english/policystatement/introduction.cfm>. In the U.K, the British Association of Applied Linguistics publishes general good practice guidelines: <http://www.baal.org.uk/goodprac.htm#6>. In Australia, an article by Chalmers (2003, <http://onlineethics.org/reseth/nbac/hchalmers.html#system>), suggests that the national statement on research will ensure a very high standard of protection for human subjects. It is beyond the scope of this chapter to discuss specific global informed consent practices, but obviously countries will have varying (and evolving) approaches to the protection of human subjects.

Ethical considerations for research involving human subjects are outlined in various publicly available international and U.S. government documents. For instance, the *Declaration of Helsinki* (World Medical Association, 2000), the *Belmont Report* (National Commission, 1979), and the Nuremberg Code (1949) are available online, and both the Office for Human Research Protections (OHRP) of the U.S. Department of Health and Human Services (DHHS) and the Office of Human Subjects Research (OHSR) of the U.S. National Institutes of Health (NIH) provide online documentation concerning ethical principles and guidelines for the protection of human subjects. (For the pertinent Web sites, please consult the reference list at the end of the book as well as the URLs provided in this chapter.) Much of the information we provide in this chapter is based on these documents, together with the online training module offered by the U.S. NIH (available online at <http://cme.nci.nih.gov>). These free and publicly available resources provide a careful review of historical events that have shaped current U.S. ethical regulations governing scientific research with human subjects. Although we summarize some of that information here, we also recommend that researchers visit some of the Web sites and online documents that have been designed to promote understanding of the processes of human subjects research. Indeed, some funding bodies require that online training modules be completed before research grants are awarded.

2.1.1. Obtaining Informed Consent From Second Language Learners

Beginning with the Nuremberg Code (1949), the notion of informed consent has become a cornerstone of ethical practice in research involving human subjects. A number of helpful sources outline in detail the essentials of informed consent—in particular, the responsibilities of the researcher—as well as typical elements of a written informed consent document. These include information provided by the U.S. Office for Human Research Protections (OHRP) in their Investigator Responsibilities and Informed Consent training module (available online at http://137.187.206.145/cbttng_ohrp/cbts/assurance/module2qset1_1.asp), which also provides details about other government guidelines and information on human subjects. The institutional review boards (IRBs) of various universities also provide guidelines for writing informed consent documents, for example, Harvard's *The Intelligent Scholar's Guide to the Use of Human Subjects in Research* (2000). IRBs are also sometimes known as *human subjects committees*, *informed consent committees*,

and *research ethics committees*, among other things. In this chapter, we consistently use the term *IRBs*.

According to the *Belmont Report* (National Commission, 1979), which was important in the development of informed consent and is discussed further later in this chapter, informed consent requires that human subjects, to the degree that they are capable, should be provided with the opportunity to choose what shall or shall not happen to them. This can occur only when at least the following three conditions are fulfilled:

1. Suppliance of sufficient information (i.e., full disclosure about the experiment by the researcher).
2. Comprehension on the part of the subject.
3. Voluntary participation, in which the subject is free from undue pressure or coercion.

Thus, the nature of consent implies voluntary agreement to participate in a study about which the potential subject has enough information and understands enough to make an informed decision. Each of these elements is discussed in more detail next.

2.1.1.1. Suppliance of Sufficient Information

What constitutes sufficient information? The answer to this question depends to some extent on what source on human subjects is consulted; different institutions (including different universities and government bodies) may have different interpretations of “complete disclosure.” However, some core elements can be identified. Among these is the idea that potential participants should be supplied with information that describes the procedures and purposes of the research, as well as the potential risks and benefits. This may sometimes be interpreted as including details such as the method by which participants will be assigned to any groups in the study (e.g., treatment groups or control group). Some institutions also agree that potential participants should receive information about whom to contact if questions arise regarding the study or the subjects’ rights. Sometimes the researchers’ contact information may be supplied on the consent form, sometimes the review board details are made available, and sometimes both are supplied. Finally, information is usually provided about the steps the researcher will take to ensure that any identifying aspects of the data will be held confidential.

These points are all applicable to second language research. For example, many review boards or human subjects committees will require that learners be informed about the procedures, purposes, and potential risks and benefits of the studies. Second language research usually does not lead to risks in the same way that some medical or psychologically based research can lead to risk. However, in research on the effect of second language instruction there might be a control group that will not receive instructional time equal to that of the experimental groups. Depending on the regulations of the body approving the research, learners might need to be informed that they could be assigned to a group that, theoretically, could benefit less than a treatment group. In the same study, if intact classes are used and group assignment is made on this basis, learners might need to be informed about this method of assignment, even if it leads them to ask questions or wish to change classes. Also, second language researchers are often required to include their contact information on informed consent documents, even if it results in their students (e.g., for teachers researching their own classrooms) calling them to discuss class work outside the experiment. Finally, confidentiality of data is important in second language research. As Duff and Early (1996) noted in their discussion of confidentiality, “[A]lthough it is common practice to change the names of research subjects, this in itself does not guarantee subject anonymity. In reports of school-based research, prominent individuals or focal subjects tend to be more vulnerable than others” (p. 21).

Also, if the researcher uses quotations in the final write-up or presentation, certain individuals may be recognizable to other researchers, perhaps because of what they say, for example in terms of their position on a topic or if the data are played audibly, perhaps by the sound of their voice. This may be less likely to apply to learners, but can certainly apply in school settings. If teachers are identified, even unintentionally, this could have ramifications for future promotions, contract renewals, or class assignments; for students, identification might have implications for how other teachers perceive them, and consequently might have an impact on their grades and letters of recommendation. Immigrant and refugee populations may also fear that sensitive information may be intentionally or inadvertently disclosed to the authorities. Neufeld, Harrison, Hughes, Spitzer, and Stewart (2001), for example, noted that “in Middle Eastern immigrant populations, individuals were distrustful of research and of the university, which they associated with the government” (p. 586).

To alleviate these concerns, researchers are advised to make it clear from the beginning that all information will remain confidential and anonymous wherever possible, and to explain the various steps that will be taken to protect the learners' anonymity (e.g., using numbers instead of names to refer to participants, not revealing identifying information, discussing the location of records and who will have access to them). In particularly sensitive situations, such as those involving refugees, second language researchers might even volunteer to check with the participants before using any potentially identifying information in transcripts, data, reports, papers, or presentations, even when numbers are assigned instead of names. In general, second language researchers need to be sensitive to concerns about anonymity. Some review boards or committees might even ask where the data are to be stored and with whom they are to be shared. For example, the use of learner corpora is growing in the field of second language research, and many corpora are freely available over the Internet. Corpora can be an excellent way to avoid duplication of effort in the extremely time-consuming practice of data elicitation, collection, and transcription. However, this practice of sharing data may lead researchers to forget that sending a colleague transcripts or data from their research may in fact not be covered under the original IRB or committee regulations and the approval that was signed, nor by the informed consent documents the learners signed. Not all universities and schools in all countries require consent and, even when they do, requirements may differ. It is therefore important to verify IRB requirements before embarking on a research project.

Checklist for Obtaining Informed Consent

Overall goal: To ensure that participants are supplied with enough information to make informed voluntary decisions about participating. This can include information about:

- The procedures and purposes of the research.
- The potential risks and benefits of the research.
- The methods by which participants will be assigned to groups and what those group assignments might entail in terms of treatment.
- Whom to contact with questions regarding the study or their rights as participants.
- The specific steps that will be taken to ensure confidentiality and anonymity; for example:

- Using numbers instead of names.
- Not revealing identifying information.
- Safeguarding the location of and access to records.

Researchers need to remember to:

- Consider the special implications that the research (and confidentiality/anonymity issues) may have for any teachers, students, or immigrant/refugee populations involved.
- Make sure any subsequent sharing of data is permissible by institutional regulations and the informed consent that was obtained from participants.

2.1.1.2. Is Withholding Information Ever Necessary?

In general, researchers are advised to provide as much information as possible to participants because failure to disclose information may constitute deception. In second language research, however, it may occasionally be necessary not to fully disclose information. We discuss this throughout the book as ‘giving away the goals of the study.’ As Rounds (1996) explained, in second language research “sometimes ... a research design requires that the researcher conceal her real interests, and perhaps use small deceptions to deal with the classic ‘observer’s paradox’” (p. 53).

For example, if the researcher is studying a teacher’s use of questions in the L2 classroom, informing the teacher about the goals of the research may bias his or her use of questions and thus lead to an unrepresentative sample of data. In this case, withholding information may be acceptable and allowed by the human subjects committee, but three conditions will often need to be met:

1. Incomplete disclosure is essential to the aims of the research.
2. No risks are undisclosed.
3. Participants will be given an opportunity to be debriefed after the study.

Researchers need to think carefully about how much deception is ethical. For example, although telling a teacher that the study is about his or her language is not fully disclosing the purposes of the study and is therefore not ideal, it might be a better route than deceiving the teacher by telling his or her that the research is focusing on how he or she uses the whiteboard. In some studies it may be appropriate to advise the participants ahead of time

that the study is about second language learning and that the exact features of it will be described to them after the study. This is a practice also used in some psychology-based research (see Baumrind, 1990, for further discussion of ethical issues and deception in applied psychology research).

In summary, then, incomplete disclosure may be acceptable in some cases and seems to be a common practice in some areas of second language research. In these cases, it may be sufficient to indicate to participants that they are being invited to participate in research for which some features will not be revealed until the research is concluded. In those instances, the purpose of the study is presented in general terms only. Based on current guidelines, such second language research can be justified only if the three conditions cited earlier are met.

2.1.1.3. Participant Comprehension in Informed Consent

In addition to supplying sufficient information to potential participants to allow them to make informed decisions, the researcher is also responsible for ensuring participant comprehension. Thus, the way in which information is conveyed might be as important as the information itself. This implies (a) that the potential participants be given the opportunity to discuss concerns and have their questions answered, and (b) that the informed consent document be provided in language understandable to them, given their age, educational background, mental capacity, and language proficiency, including literacy (National Commission, 1979).

Second Language Learners. Clearly, these guidelines are important when we consider the case of second language learners, who are often asked to read and sign consent forms that are not written in their first language. One general suggestion from the OHRP is to write the consent documents so that they are understandable to people who have not graduated from high school. However, for low-proficiency language learners, it may be necessary to provide a translation of the consent document in the learners' first language. Alternatively, many sources suggest that the document be presented orally in the first language, along with a short form of the consent document attesting to the oral presentation in the first language and containing a summary of the information presented orally. The use of research assistants who speak the participants' L1 may be especially valuable in this respect. When the use of the L1 is not possible or practicable (e.g., when the researcher is studying a large population

with multiple L1s), the document should be written (or orally provided) in the L2 at a level understandable to the potential subjects (OHRP Training, n.d.). In the box that follows we present some guidelines that are based on government recommendations.

Consent Form Guidelines for Nonnative Speakers

Wherever possible, *informed consent information* should be presented in a language that is understandable to the subject. For example, for ESL speakers, that should be their native language or a language in which they are highly proficient.

- *Best option:* The written consent form is translated into the native language(s) of the learner unless the learners are clearly proficient enough to understand the form. Both the English consent forms and the translations should be approved by the review board or human subjects committees. Translated forms are often presented on the reverse side of the sheet
- *Alternative option:* A translator can explain the consent form to the learner. This option might be best if the learner is not literate in his or her L1, if the L1 has no written form, or if a written translation has proved very difficult to obtain. If the consent form is explained orally, researchers need: (a) a bilingual translator (who also serves as a witness); (b) a short consent form in a language understandable to the speaker, and (c) a review board- or human subjects committee-approved L1 version of the consent form. The following process can be followed: (a) The translator explains the consent form to the participant; (b) both the participant and the translator sign the short form; and (c) the researcher and the translator sign the English version.

In appendix A, we provide an example of an abbreviated informed consent form that might be used, together with a translator, for learners whose native language is not English. However, we urge researchers to check with their own institutional review board for guidelines and sample consent forms. Figure 2.1 provides an example of a complete informed consent form that might be used with learners whose native language is not English, but who are deemed proficient enough to understand the form.

Child Second Language Learners. When collecting data from children for second language research purposes there are several important things to consider. As Thompson and Jackson (1998) noted, “[S]econd language researchers must keep in mind that children cannot be treated just like adults as research subjects. Because their capabilities, perspectives, and needs are differ-

Consent to Participate in Research

Project Name _____

L2 learners' performance on grammaticality judgment, oral production, and listening tasks

Investigator _____ Telephone _____

E-mail _____

Sponsor

None (*The University Institutional Review Board has given approval for this research project. For information on your rights as a research subject, contact _____*)

Introduction

You are invited to consider participating in this research study. We will be comparing the performance of EFL learners on three different tasks: a speaking activity, a written activity, and a listening activity. This form will describe the purpose and nature of the study and your rights as a participant in the study. The decision to participate or not is yours. If you decide to participate, please sign and date the last line of this form.

Explanation of the study

We will be looking at the kind of language you use when you do three different kinds of activities: a speaking activity, a writing activity, and a listening activity. About 40 students enrolled in _____ will participate in this study. As part of the study, you will meet with the researcher for an oral interview. At the same time, you will do the writing activity and then the listening activity. All three tasks will take about 30 minutes to complete. A tape recorder will be used to record what you are saying during the speaking activities.

Confidentiality

All of the information collected will be confidential and will only be used for research purposes. This means that your identity will be anonymous; in other words, no one besides the researcher will know your name. Whenever data from this study are published, your name will not be used. The data will be stored on a computer, and only the researcher will have access to it.

Your participation

Participating in this study is strictly voluntary. Your decision to participate will in no way affect your grade. If at any point you change your mind and no longer want to participate, you can tell your teacher. You will not be paid for participating in this study. If you have any questions about the research, you can contact _____ by telephone at _____, by e-mail _____, or in person at the office in _____.

Investigator's statement

I have fully explained this study to the student. I have discussed the activities and have answered all of the questions that the student asked.

Signature of investigator _____ Date _____

Learner's consent

I have read the information provided in this Informed Consent Form. All my questions were answered to my satisfaction. I voluntarily agree to participate in this study.

Your signature _____ Date _____

FIG. 2.1. Sample consent form for an experimental study.

ent, children approach the research context uniquely and encounter a different constellation of research risks and benefits from their participation” (p. 223).

Because of this, the researcher needs to explain the research in language that is comprehensible and meaningful to the child; in addition, the researcher needs to inform the child’s parents about the nature and consequences of the study, as well as obtain a signed consent form from the parent. Researchers will also need to assure school boards and parents that the procedures used in the research will not negatively impact the second language learning process or pose any more than a minimal risk to the physical and psychological well-being of the child. As defined by the U.S. Department of Health and Human Services (2003), *minimal risk* is a risk of harm not greater than that “ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests” (p. 113). With respect to potentially vulnerable research subjects such as children, the Institutional Review Board at the University of Pennsylvania stipulated the following:

[R]esearch in children requires that the IRB carefully consider consent, beneficence, and justice Children may be subjects of research only if informed consent is obtained from the parents or legal guardian. Children over the age of 7 must agree to participate in the research and provide written assent and separate assent forms should be provided based on reasonable age ranges for comprehension. (n.d., pp. 3–4, <http://www.upenn.edu/regulatoryaffairs/human/SOP/SC-501.pdf>)

2.1.1.4. Voluntary Participation and Informed Consent

Invitations to participate in research must involve neither threats of harm nor offers of inappropriate rewards. Undue influence may be exercised, even unwittingly, where persons of authority urge or suggest a particular course of action.

For second language research, care must be taken, for example, when classroom teachers invite their students to participate in a study. Even when it is clear that there will be no extra points or higher grades for participation and no penalty for declining to participate, the simple fact that the teacher is the one requesting the students’ participation may constitute undue influence. For this reason, it can often be helpful for the teacher to distance him- or herself from the process and leave the room while a third party explains

the research and distributes the forms. Even when the teacher is the researcher, this course of action may be preferable so as to avoid potentially influencing the students. Some universities may go as far as prohibiting researchers from carrying out research on their own students, although it should be noted that even those universities that have a stated policy along these lines may allow some flexibility. For example, Indiana University's institutional review board guidelines in this respect are as follows:

The Committee has long taken the position that teachers should not use their own students as subjects in their research if it can be avoided. This general policy is in accord with that of other institutional review boards. The Committee recognizes, however, that in some research situations, use of one's own students is integral to the research. This is particularly true of research into teaching methods, curricula and other areas related to the scholarship of teaching and learning. (1999–2003, <http://www.indiana.edu/%7Eeresrisk/stusub.html>)

Finally, we emphasize the need to check with local authorities regarding standard procedures. For example, in some countries there is suspicion concerning consent forms, the idea being: "Why is a consent form needed? It might be an indication that something bad can happen." In other words, consent is a common part of the research process in some parts of the world, but not in others. This can create a dilemma when researchers must have signed consent given the regulations of their universities, but are conducting research in an environment where signed consent is looked on with suspicion. It is important to verify the research climate in the setting where the research will be conducted. If a conflict is likely to exist, it must be dealt with before beginning data collection to avoid serious problems after data have been collected.

2.1.1.5. The Informed Consent Document

Guidelines for creating informed consent documents for research appear in government publications (e.g., OHRP) as well as in the IRB guidelines of various universities. Appendixes A, B, and C of this volume contain several examples of informed consent documents relevant to second language research for the reader to consult and compare. From these documents it should be apparent that although there are differences across subject types, the documents tend to be highly similar. The following checklist may be helpful in drawing up informed consent documents:

- Does your informed consent document explain the general procedure of the research?
- Does it explain the purpose of the study?
- Does it include information about the risks (and benefits) to which participants may be exposed, as well as steps you will take to minimize any risks?
- Does it provide information about how the participants' identities will be kept anonymous and confidential as much as possible?
- Does it provide contact information so that the participants can contact you or you and/or your human subjects committee if they have questions or concerns?
- Does it make it clear that participation is purely voluntary and that the participants have the right to withdraw at any time?
- Have you checked to make sure the document does not contain any language suggesting that the participants are giving up any rights or are releasing the investigators from any liability?
- Is your document written in language understandable to the participants (in the participants' L1 or, alternatively, in basic English that avoids technical jargon)?
- Have you considered how you will provide the participants with ample time to review the document before making their decision?
- If the potential participants do agree to participate, have you checked to make sure the documents are dated on the same day they are signed?
- Have you considered that multiple consents may need to be obtained for one study; for example, from parents, teachers, child learners, school administrators, and so on?
- Finally, have you given a copy of the signed consent form to all those who agreed to participate and kept an original signed copy for your own records?

2.1.2. History of Institutional Review of Human Subjects Research, Compliance, and Problem Solving

2.1.2.1. Purpose of Reviews and IRB Responsibilities

As noted at the beginning of this chapter, some second language researchers wonder why it is necessary to adhere to human subjects guidelines for research areas in which no risk is involved (e.g., judging sentences

as to their correctness). In this section, we provide a historical overview of research with human subjects in order to provide context for today's regulations. In the United States, in 1974, the National Research Act, which created the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, also required the establishment of Institutional Review Boards (IRBs) to review all research funded by the Department of Health and Human Services. Requirements for IRBs have been periodically revised and widely adopted. Thus, most major universities in the United States have institutional review boards that review all research involving human subjects, including all second language research. As noted previously, not all are termed *IRBs*; the term *human subjects committees* is also often used interchangeably with *IRBs*.

IRB reviews are designed to ensure the protection of human research subjects. In general, the job of the IRB is to ascertain that the investigator is in compliance with federal standards for ethics and is taking adequate steps to protect the rights and well-being of research participants. This includes equitable selection of subjects, adequate communication of information and risks in informed consent documents, and clear statements about the confidentiality of data. It is also the responsibility of the IRB to investigate any alleged or suspected violations of protocols. IRBs usually have the right to suspend or require modifications to research protocols if they deem such action necessary.

The IRB review may also entail, to some degree, assurance of the scientific merit of the research, although U.S. federal regulations leave IRBs without clear direction on this point. According to the OHRP Guidebook for IRBs (n.d., http://ohrp.osophs.dhhs.gov/irb/irb_guidebook.htm), many IRBs appear to take the following approach: If the investigator is seeking external funding for the research, review of the scientific quality of the research involved is left to the agency; if no such funding is being sought, the IRB may take more care in reviewing the research design.

In relation to second language research, it is important to note that full IRB reviews can be time consuming when they are carried out. However, second language (and other linguistically oriented) research often qualifies for either exempt status (no IRB review required) or expedited review, depending on the policy of the institution's IRB. For example, observational studies (e.g., the cataloguing of adult behavior in public places, with no manipulation of the environment) are generally exempt from IRB review because they are considered low risk. Expedited reviews also apply to certain kinds of low-risk research, such as minor changes in previously approved research during the

authorized approval period. Sometimes classroom observational research is considered low risk, as is the examination of data collected from regular classroom tests. In an expedited review, all of the requirements for IRB approval of research are in effect. However, in expedited reviews, approval may be granted by the chair of the IRB alone, or by one or more members of the IRB, instead of requiring a full review by the entire board. Expedited reviews are processed more quickly, but the full protocol for an application for a complete review is usually required in addition to the application for expedited review. Examples of application forms for exempt status and full, continuing, and expedited review can be found in appendixes D through G. As in all instances of human subjects research, researchers are urged to check with their local IRB for specific requirements.

2.1.2.2. Why Guidelines to Protect Human Subjects Were Developed

Various troubling practices of the past have raised questions about the ethical treatment of human subjects in research. Although the detail that follows on medical experimentation may seem at first glance to be some distance from second language research, it is important to understand that the various principles developed out of the statements and reports have direct consequences for guidelines involved in carrying out all research involving human subjects. This includes research involving second language learners.

Perhaps the most notorious violation of human rights, with a substantial number of people affected and a significantly high degree of harm caused, is the case of the Nazi “medical experiments” carried out in concentration camps during World War II. In the Nuremberg Military Tribunals, at which 23 physicians and administrators were indicted for their participation in these experiments, the judges provided a section in their written verdict entitled “Permissible Medical Experiments,” which has become known as the Nuremberg Code (1949). A number of aspects of this important code inform current research practices.

Another disturbing example of the prolonged and deliberate violation of the rights of a vulnerable population is the well-known U.S. Tuskegee Syphilis Study (1930–1972). The course of untreated syphilis was studied in a group of 399 infected Black airmen. These men were not only recruited for the study without their informed consent, but were also misinformed because researchers implied that they were receiving treatment. Even when penicillin

was confirmed as an effective treatment for syphilis in the 1940s, the men were neither informed nor treated. Press coverage of this study began to appear in 1972. This, together with other biomedical abuses, led to public outrage in the United States and eventually to the U.S. Congress establishing a permanent body to regulate all federally supported research involving human subjects. This body is now known as the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Many other violations of the rights of human subjects can be found in the literature. Collectively, these events helped to shape the course of the development of an ethical code for research practices.

2.1.2.3. Development of Research Codes of Ethics

The Nuremberg Code (1949), which has served as the basis for many later ethical codes governing research involving human subjects, outlines 10 directives for human experimentation, the first of which is that voluntary, fully informed consent to participate in experiments is essential. Also, the experiment should be necessary and fruitful for the good of society, should avoid unnecessary physical and mental suffering and injury, and should be based on prior animal research where applicable. Human subjects should be protected against possibilities of injury or death and be allowed to withdraw from the experiment; moreover, the degree of risk should not outweigh the experiment's humanitarian importance. This code was, in effect, the first international code of research ethics.

The *Declaration of Helsinki*, first adopted in 1964, was developed by the World Medical Association (2000, <http://www.wma.net/e/policy/b3.htm>) to serve as a statement of ethical principles providing guidance to physicians and other investigators in medical research involving human subjects. It was the medical community's first major attempt at self-regulation, and it is periodically revised and updated. Like the Nuremberg Code, it emphasizes the importance of the voluntary consent and protection of the subject, the subject's freedom to withdraw from the research, and the idea that potential benefits should outweigh foreseeable risks. It also requires special protections for vulnerable research populations.

The *Belmont Report* (National Commission, 1979) in the United States, based on the work of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, identified three major ethical principles that should guide individual and

institutional considerations in human research: respect for persons, beneficence, and justice. These basic principles appear in the NIH online training module mentioned earlier. *Respect for persons* entails that all people should be treated as autonomous agents and that those with diminished autonomy are entitled to special protections. *Beneficence* involves respecting people's decisions and protecting them from harm, as well as making efforts to ensure their well-being. It has been taken to imply two general rules: Do no harm, and maximize possible benefits and minimize possible risks. *Justice* means fairness in the distribution of the risks and benefits of research.

2.1.2.4. Preparing a Protocol for the IRB

A protocol is essentially an application to an IRB for approval to carry out research. IRBs generally provide an application form that solicits the information they need for their review. This often includes a template for an informed consent document. As mentioned earlier, examples of IRB application forms appear in appendixes D through G, and examples of informed consent documents appear in appendixes A through C.

The protocol to be submitted generally requires the following sections, although different institutions may name or organize these sections differently. In the *précis* (also known as the abstract), the researcher provides a short overview of the study's objectives, population, design, and potential results of interest. In the introduction, background information is provided, along with a review of the relevant literature. In the objectives section, the researcher states the objectives of the study and, whenever possible, the hypotheses. Next, in the study design and methods section, the researcher describes how the participants will be chosen and how treatment is to be administered. In the data analysis section, the researcher explains how the outcomes will be measured and statistically analyzed. After this, in the human subjects protection section, the researcher usually provides:

1. Information about any strategies or procedures for recruitment (including advertising, if applicable).
2. A description of the potential benefits that may reasonably be expected from the research.
3. An explanation of the potential risks (physical, psychological, social, legal, or other), their likelihood, and what steps the researcher will take to minimize those risks.

4. A description of the consent procedures to be followed.

A section on the qualifications of the researcher to carry out the study is also often required. Finally, in the references section, the researcher provides a list of studies that were cited in the protocol (OHSR, Information Sheet #5, 2000).

General Use of Protocols in Research. In addition to being an essential step in obtaining IRB approval, protocols are also useful in that they can be used to provide a ‘roadmap’ for researchers. Because method and design in second language research can often be complex, a detailed protocol “helps the researcher to anticipate problems in advance while also acting as a checklist for the many variables and factors the researcher needs to consider and balance while carrying out the procedure” (Gass & Mackey, 2000, p. 57). A detailed protocol also ensures that if multiple individuals are responsible for administering a test or collecting data, for example, they will do so in a uniform fashion. Preparation of a detailed research protocol often goes hand in hand with a pilot study (see chap. 3) as part of the preparatory steps in carrying out second language research. Thus, even if researchers feel constrained by the requirements to make applications to IRBs, human subjects committees, or ethics boards, they should feel encouraged that the preparation of a protocol can be helpful in thinking through and planning out the steps involved in the research.

2.2. CONCLUSION

In this chapter, we reviewed ethical issues in the use of human subjects in second language research, including informed consent, IRBs, and protocol preparation. Knowledge about how to ensure that participants are adequately informed about research and their rights as participants will foster confidence between the research community and the public, whereas expertise in writing research protocols and knowledge about them can lead to legal confidence for IRBs and researchers. Finally, awareness of ethical issues is likely to lead to more thoughtful and ethical research practices, which are clearly of benefit to everyone. Thus, gaining a broad view of the issues reviewed here is sure to benefit researchers, both novice and experienced. In chapter 3 we turn to a discussion of eliciting data.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Why is informed consent important for research involving second language learners?

2. When planning a classroom study of 30 ESL learners aged 11–12 years old, discuss how you would resolve the following issues:
 - a. The children are not required by the school to give their consent, but you wish to supplement your parental consent form by giving the children the opportunity to opt out. Indicate the principles by which you would modify the parental consent form for the children.
 - b. Two children and their parents decline to consent. What steps can you take to avoid collecting data from these children on your videotapes? (Obviously, you will not *use* any data collected in this way, but there are also steps you might take to avoid collecting them.)
 - c. Halfway through the study, one child's parents withdraw him, and a second child says she will be absent for the next five classes. How do you deal with each data set?
 - d. The potential participants for your research tend to be suspicious about researchers and are hesitant to participate for fear that sensitive information may be revealed about them upon completion of the research. What sort of information could you provide to reassure them?
3. What is an IRB protocol?
4. Explain the difference between an expedited and a full IRB review.
5. Why is it important to provide as much detail as possible in research protocols?

Common Data Collection Measures

Data collection in second language research is multidimensional. In this chapter we provide details about many of the measures that second language researchers commonly use to collect data. We have divided the chapter into sections according to research area, providing a sample of methods used in each area. We include methods used with formal models, processing-based research, interaction-based research, research on strategies and cognitive processes, sociolinguistic- and pragmatics-based research, and questionnaire- and survey-based research. Finally, we briefly discuss existing databases in the second language research field. We begin with a discussion of the importance of pilot testing.

3.1. PILOT TESTING

A pilot study is generally considered to be a small-scale trial of the proposed procedures, materials, and methods, and sometimes also includes coding sheets and analytic choices. The point of carrying out a pilot study is to test—often to revise—and then finalize the materials and the methods. Pilot testing is carried out to uncover any problems, and to address them before the main study is carried out. A pilot study is an important means of assessing the feasibility and usefulness of the data collection methods and making any necessary revisions before they are used with the research participants.

In general, it is crucial for researchers to allocate time for conducting pilot tests. Although it might seem that careful advance planning and preparation might allow the researcher to skip pilot testing, it is in fact critical, because it can reveal subtle flaws in the design or implementation of the study that may not be readily apparent from the research plan itself. As Gass

and Mackey (2000) explained, pilot testing “can help avoid costly and time-consuming problems during the data collection procedure ... [as well as] the loss of valuable, potentially useful, and often irreplaceable data” (p. 57). We return to the importance of pilot testing in several of the following chapters as we consider issues such as data coding and classroom research.

Sometimes pilot studies result in data that might be useable for the main study. Some researchers choose to seek permission from their human subjects committees or institutional review boards to carry out an experiment in such a way that if they do not encounter problems with their pilot testing, they can use the data for their main study as long as it is collected in exactly the same way. However, not all institutions will give permission for this, and many do not have a process for the retroactive use of data. It is worthwhile to investigate these issues while also keeping in mind that it is a rare pilot study that does not result in some sort of revision of materials or methods.

3.2. THE SIGNIFICANCE OF DATA COLLECTION MEASURES

Findings in second language research are highly dependent on the data collection (often known as *data elicitation*) measures used. One goal of research is to uncover information about learner behavior or learner knowledge independent of the context of data collection. There is no single prescribed elicitation measure, nor is there a “right” or “wrong” elicitation measure, although many research paradigms have common measures associated with them. Saying that there are numerous elicitation measures that have been used in second language research does not, however, imply that measures are random or that one is as good as another. The choice of one measure over another is highly dependent on the research question asked and may also be related to the theoretical framework within which research is conducted. It is the purpose of this chapter to present some common elicitation measures. Although we have organized this chapter by research paradigms, the reader should be aware that this is in some sense for the sake of convenience and that there is often crossover from method to paradigm. For example, stimulated recall, which is discussed in this chapter as part of strategies-based research, has also recently come to be used for research on noticing and interaction. The measures that we have chosen to describe here do not represent an exhaustive list; in fact, a complete list would be impossible because elicitation measures are constrained only by the limits of researchers’ imaginations.¹

¹For further information on elicitation methods and particularly on interlanguage analyses, see Gass and Selinker (2001, chap. 2).

We mentioned previously that research questions to a certain extent dictate a particular method. Let us consider what this means by looking at some hypothetical examples.

3.2.1. Syntax: Japanese Passives

After an extensive literature review and after years of teaching Japanese, you recognize that English-speaking learners of Japanese have great difficulty with passives. You do some reading and realize that there are theoretical reasons for this, so you decide to investigate. The task in front of you is to gather data from learners of Japanese to determine exactly what faulty generalizations they may be making. In other words, what are the learner-language forms that are being used at various stages of Japanese proficiency? Once you have determined what the data are that you need to elicit (i.e., samples of Japanese passives), your next task is to determine how to elicit appropriate data.

Your first thought is to have learners describe pictures that depict various actions (e.g., a man being hit by a ball, a dog being kissed by a boy). Unfortunately, the learners, who are experts at avoidance, produce very few examples of the structure in question. You then modify the task and tell the learners to start with the patient; that is, with the object of the action. You even point to “a man” and to “a dog.” This doesn’t work very well because the learners do not do what is being asked; they still begin with “ball” and “boy.” You are thus left with the question: Did they not produce the requisite passive because (a) they don’t have the linguistic knowledge, (b) the active sentence is easier to formulate, or (c) they didn’t understand how to carry out the task? There are too many possibilities as you attempt to interpret the data. Only the first interpretation will help you in dealing with your research questions. It is therefore necessary to question the value of this research elicitation method.

You then realize that you have to “force” the issue and make learners behave in a way that allows you to be confident that you are obtaining information that reflects their actual knowledge about passives. There are a few ways that this is commonly done in second language research. One way is to use what are known as *acceptability judgments* (sec. 3.3.1), by which a list of grammatical and ungrammatical sentences is presented to learners who are then asked to indicate whether they consider them acceptable Japanese sentences or not. This is followed by a request to the learners to correct those sentences they have judged to be incorrect. This method ensures that

at least the sample of sentences in which you are interested (passives) is the target of investigation. Another way to gather information about passives is through “elicited imitation” (see sec. 3.3.2.). In this method, sentences are read to the learners (usually tape recorded to ensure that everyone hears the same sentences at the identical rate and with identical intonation), and the learners are asked to repeat them. As with acceptability judgments, the researcher can control all aspects of the sample sentences. A third possibility for eliciting information about passives is known as *truth-value judgments* (see sec. 3.3.4). These might be particularly useful in the case of Japanese passives, because some of the differences involve subtle meaning differences. With truth-value judgments, learners are given short contextualized passages with relevant sentences embedded in them. Following each passage is a question that ascertains whether or not a learner can correctly interpret the meaning of the sentence.

Thus, the investigation of a particular grammatical structure offers a number of possibilities for data elicitation measures, the choice of which will depend on the questions being asked (e.g., truth-value for subtle meaning differences, or acceptability judgments or elicited imitation if one wants to gather information about grammatical knowledge). In any event, the specific research question can be used to narrow the choice of data collection measures.

3.2.2. Interaction Research

Suppose your research interest is to find out whether recasts or negotiation will lead to faster development of relative clauses in a second language.²

²A recast is usually defined as a nontargetlike NNS utterance that has been rephrased in a more targetlike way while maintaining the original meaning. Following is an example (constructed).

NNS: I have three dog my picture.

NS: You have three dogs in your picture? ← Recast

Negotiation is usually defined as two or more speakers working together to achieve message comprehensibility. In the following constructed example, the word *bird* is being negotiated.

NNS: I see bud my picture?

NS: You see what in your picture?

NNS: I see bud my picture?

NS: Bud?

NNS: Yes

NS: Do you mean bird?

NNS: Ah, bird, yes.

You find groups of English learners of Italian who are at four different stages of development in their knowledge of relative clauses. For each group, half of them will serve in your “recast” group and the other half in your “negotiation” group. You first give a pretest to ensure comparability of groups. Everyone does a picture-description task in which correction is provided by a native speaker of Italian, either in the form of recasts or negotiation. At the end of the session, you give each group a posttest on relative clauses and then a second posttest 3 weeks later. You find that there are no differences between the groups. When you go back to analyze the actual transcripts of the sessions, however, you realize that your findings are probably due to a lack of relative clause examples in the data. This example illustrates how important it is in task-based research to first ascertain whether or not your tasks will, in fact, elicit your targeted grammatical structures and provide opportunities for interactional feedback. Because relative clauses are frequently used to separate one object or person from others (e.g., *the boy who is wearing a red hat is my brother*—in other words, not the boy who is wearing the green hat), you would want to make sure that your task is consistent with the normal function of relative clauses and involves having learners identify an object or person from among others. In sum, you need to ensure (through piloting) that each elicitation measure yields the kind of data that will be useful as you address your research question.

3.2.3. Pragmatics Research

Assume that you want to conduct research on pragmatic problems that a particular group of students might have (e.g., English speakers learning Chinese). You have further limited your research to interactions between English learners of Mandarin and their Mandarin-speaking professors. You obtain permission to observe interactions between these two groups of people in order to determine what sorts of pragmatic errors may occur, but after 5 days of observations you have little in the way of consistent results. Why might that be the case? One reason might be that you have not narrowed down your research question sufficiently (see chap. 1 for more information on research questions). For example, you might be looking for too many pragmatic areas (e.g., complaining, apologizing, requesting, inviting) rather than constraining the data. A second reason is that waiting for a language event to occur may be dependent on luck. You might end up with some interesting examples that could be fodder for insightful qualitative analyses, but if you are looking for sufficient examples to be able to make

quantitative generalizations, you may need to force the issue through discourse completion tests (sec. 3.7.3) or well-designed role plays (sec. 3.7.4). In their simplest form, discourse completion tests present to learners a context in which a response is required. Role plays involve “acting” situations and are also useful for establishing specific contexts. Data could be set up so that a learner and another person are sitting at a table when the second person accidentally knocks over the water. What the learner actually says to accept the apology could then be recorded.

As can be seen through the examples provided in section 3.2, the research questions will help guide you to more or less appropriate elicitation measures. In the remainder of this chapter, we present some common methods used in a variety of research paradigms. As we noted earlier, this list is neither exclusive nor exhaustive and may even cross categories. Some might quibble with our selection and sequencing, but limits had to be imposed.

3.3. RESEARCHING FORMAL MODELS OF LANGUAGE

The first paradigm that we consider in this chapter on elicitation methods is Universal Grammar (UG), arguably the most common paradigm within the general category of formal models. The UG approach to second language acquisition begins from the perspective of learnability. The assumption is that there are innate universal language properties, and innateness is invoked to explain the uniformly successful and speedy acquisition of first language by normal children despite ‘incomplete’ (‘impoverished’) input. In UG theory, universal principles form part of the mental representation of language, and properties of the human mind are what make language universals the way they are. As Chomsky (1997) noted, “The theory of a particular language is its *grammar*. The theory of languages and the expressions they generate is *Universal Grammar* (UG); UG is a theory of the initial state S_0 of the relevant component of the language faculty” (p. 167). The assumption that UG is the guiding force of child language acquisition has long been maintained by some researchers. The question for second language research concerns the extent to which UG may be available to second language learners. Or, to put it differently, to what extent does UG constrain the kinds of second language grammars with which learners can come up?

3.3.1. Acceptability Judgments

The theory underlying UG assumes that language is constrained by a set of abstract principles that characterize core grammars of all natural lan-

guages. In addition to principles that are invariant (i.e., characteristic of all languages) are parameters that vary across languages. UG-based second language research seeks to determine the extent to which second language learners have the same abstract representations as do native speakers. One needs to determine, then, what learners believe to be grammatical in the language being learned and what they consider ungrammatical. Importantly, second language input alone does not provide learners with this information (see Gass & Selinker, 2001, chap. 7, for further details). Because UG researchers need to understand the nature of abstract representations and because abstractness is only inferable from surface phenomena, researchers are often in a position of needing to force learners into stating what is possible and what is not possible in their second language. Acceptability judgments, a common elicitation tool in linguistics, are often used for this purpose.³

In an acceptability judgment task, learners are asked whether a particular sentence is acceptable in the second language. As mentioned earlier, some sort of forced elicitation may be necessary in cases in which researchers want to investigate a particular grammatical structure because otherwise they might have to wait a considerable amount of time for enough instances to occur in natural production to draw reasonable conclusions, and with time, of course, changes can occur in second language grammars. Moreover, part of understanding what someone knows about language is understanding what they include in their grammar and what they exclude. This cannot be inferred from natural production alone, as the following example demonstrates:

Research question: Do learners know the third-person singular –s in English?

Typical production: The man walks down the street.

If learners consistently produce sentences such as this, we might be able to conclude that they have knowledge of third-person singular. However, that is a conclusion based on insufficient evidence. It is a valid conclusion only if we know what the learners exclude—that is, if we know that they rule out as a possible English sentence **The boys walks down the street*. In

³It is often the case that the term *acceptability* judgment is used interchangeably with *grammaticality* judgment. This is technically not correct. We make inferences about grammaticality based on judgments of acceptability, but because grammar is abstract we do not ask about it directly.

other words, do they recognize that *-s* is limited to third-person singular as opposed to being a generalized present-tense marker?

In addition, if learners do not use a form at all, we cannot assume that they cannot use the form unless they consistently do not use it in a required context.

Over the years, there have been numerous challenges and controversies surrounding the use of acceptability judgments. Among them are questions relating to just what individuals are doing when they make judgments. In other words, what sort of knowledge is being tapped? For example, particularly with second language learners, we can ask whether their responses to sentences are truly representative of their knowledge of the second language or whether they are trying to remember what some teacher said should be the case. In fact, learners will often say that a particular sentence is not possible (as in **The boys walks down the street*), but will still continue to utter sentences like this. How does one reconcile their apparent “knowledge” with their practice? It is important to remember that native speaker judgments are tapping a system that the individual has command over. This is not the case with nonnative speakers, who are being asked about the second language while inferences are being made about another system: their interlanguage.

The general procedure for using acceptability judgments is to give a list of target language sentences to be judged and to ask for corrections of any sentences judged incorrect. There are numerous practical considerations when dealing with collecting judgment data, some of which we address next (see also Schütze, 1996).

3.3.1.1. *Materials*

Order of Presentation. Counterbalance so that order of presentation does not affect results (see Gass, 1994, for a study in which each participant received a different sentence ordering).

Number of Sentences. Balance grammatical and ungrammatical sentences. Judging sentences can become tiresome and judgments may become unreliable. One way of countering this is to make sure that participants have different orders (see previous point). In this way, sentences at the end of one order for some individuals will be in a different place in the sequence of sentences for others. This reduces the possible interpretation that the ordering affected the results. Another way is to limit the number of sentences given. Although studies have been carried out with as many as

101 (Hawkins & Chan, 1997) and 282 (Johnson & Newport, 1989), we recommend no more than approximately 50 sentences. If more are necessary, it would be advisable to follow Johnson and Newport's practice of presenting them in blocks with a break in between and letting the participants know that they can rest if they are tired.

It is also important to make sure that there are sufficient numbers of filler or distractor sentences so that participants in a study cannot easily guess what the study is about. As we see in chapter 4, this would be a threat to internal validity. If a researcher is investigating a number of structures in one study, it may be possible for the structures to serve as distractors for one another.

Timing. With tests given orally or on the computer, the timing between sentences can be controlled. Questions about how much time should elapse have not been seriously addressed in the literature, but could potentially affect the results (see Murphy, 1997). Another related question to consider is whether participants should be allowed to go back and change their answers. Given that one is attempting to gain knowledge about a learner's "grammar" and not about formal rule knowledge, it is advisable to get "quick" responses without a great deal of thinking time. With an orally administered test or with a computer-based test, this is relatively easy to control; with a paper and pencil test, one could, for example, give everyone a pen with nonerasable ink so that answers cannot be changed.

Context. So that the imagination of a context does not enter into the picture, sentences can be embedded in a situational context (see later discussion of truth-value judgments).

Comparisons. At times, one wants to have sophisticated judgments of acceptability involving subtleties in language. In such instances, it might be advisable to ask participants to judge sentences as they relate to one another (e.g., *We didn't dare to answer him back* versus *We didn't dare answer him back*—which is better?). This technique is not often used in second language research (see also sec. 3.3.3 on magnitude estimation).

Modality. Are sentences given orally, on paper, or on a computer? Examples of each type abound, although the paper and pencil task is probably the most common, with computer testing increasing in recent years. The least common is oral administration (although see Ionin & Wexler, 2002, for an example of this).⁴

⁴See review article by Murphy (1997) on the effects of modality on outcomes.

Pictures. Although not actual judgments of acceptability, a slight variation on acceptability involves the use of pictures. In these situations, learners might have to match up a sentence with a picture (see Bley-Vroman & Joo, 2001, for examples) or provide a judgment about a sentence in relation to an accompanying picture (as in Montrul, 2001).⁵ Information from these tasks may lead to inferences regarding grammaticality.

3.3.1.2. Procedures

Corrections. Certain assumptions that are often made with native speaker judgments cannot always be made with nonnative speaker judgments. For instance, one can reasonably assume with native speakers that the area targeted by the researcher will also be the one targeted by the native speaker judge. With nonnative speakers, this is not the case, because their grammars can be nonnativelike in many ways. Thus, in order to ensure that the research focus is the same as the nonnative speakers' focus, it is necessary to ask learners to correct those sentences they judge to be unacceptable. For example, given the sentence **She knows the woman whom is the sister of my friend*, learners might judge this as incorrect. But without correction, we do not know what they think is incorrect about it. If our target is relative pronouns (*whom*) and they change the sentence to *She knows the woman whom is my sister's friend*, we can make the assumption that they believe that *whom* is correct. This is an important consideration when contemplating scoring.

When should we ask participants to make corrections? This is generally done immediately following the judgment. In other words, participants judge a sentence as correct or incorrect and make corrections to the sentences that were judged incorrect. The instructions (discussed next) should generally include information about corrections as part of what participants are expected to do. In Gass and Alvarez-Torres (2005), judgments were collected on a computer. At the end of the session, the computer printed out all sentences for which the participant had given a judgment of incorrect. At that point the participant was asked to make corrections. This has the potential of reducing the possibility that participants are responding "not good" due to uncertainty about correction.

Instructions. The idea of rating sentences for grammaticality/acceptability is novel to many participants. Participants often confuse

⁵See Juffs (2001) for a discussion of the necessity of using pictures for some types of linguistic information and for the drawbacks of using still pictures.

“making sense” with judgments of grammaticality. For example, the sentence *The table bit the dog* is grammatical in the pure sense; there is a noun phrase and a transitive verb followed by another noun phrase. However, because the first noun is inanimate and the second is animate, the sentence doesn’t make sense; tables can’t bite dogs. Instructions, therefore, and the examples provided need to be carefully crafted. For example, Ionin and Wexler (2002) provided explanation orally: “The investigator talked the practice items over with the child and ensured that the child was responding to the grammaticality and not to the meaning of the sentences” (p. 121).

Birdsong (1989) gave examples of some unsuccessful instructions:

- Do the following sentences sound right?
This does not eliminate the problem of confounding grammaticality with sentences that are semantically anomalous.
- Tell me if for you this makes a sentence.
A “sentence” such as “When will your grandmother arrive?” may be rejected because it is a *question* and not a *sentence*. (pp. 114–115)

One of the most thorough sets of instructions comes from Bley-Vroman, Felix, and Ioup (1988) in their investigation of accessibility to UG with a specific focus on *wh*-movement by Korean learners of English:

Sentence Intuitions

Speakers of a language seem to develop a “feel” for what is a possible sentence, even in the many cases where they have never been taught any particular rule. For example, in Korean you may feel that sentences 1–3 below sound like possible Korean sentences, while sentence 4 doesn’t. [The sentences below were actually presented in Korean.]

1. Young Hee’s eyes are big.
2. Young Hee has big eyes.
3. Young Hee’s book is big.
4. Young Hee has a big book.

Although sentences 2 and 4 are of the same structure, one can judge without depending on any rule that sentence 4 is impossible in Korean.

Likewise, in English, you might feel that the first sentence below sounds like it is a possible English sentence, while the second one does not.

1. John is likely to win the race.
2. John is probably to win the race.

On the following pages is a list of sentences. We want you to tell us for each one whether you think it sounds possible in English. Even native speakers have different intuitions about what is possible. Therefore, these sentences cannot serve the purpose of establishing one's level of proficiency in English. We want you to concentrate on how you feel about these sentences.

For the following sentences please tell us whether you feel they sound like *possible* sentences of English for you, or whether they sound like *impossible* English sentences for you. Perhaps you have no clear feeling for whether they are possible or not. In this case mark *not sure*.

Read each sentence carefully before you answer. Concentrate on the structure of the sentence. Ignore any problems with spelling, punctuation, etc. Please mark only one answer for each sentence. Make sure you have answered all 32 questions. (Bley-Vroman et al., 1988, p. 32)

It is often beneficial if instructions can be translated into the native languages of the participants. The Bley-Vroman et al. instructions would, of course, need to be modified according to the specific circumstances.

Scoring. Scoring will depend on how the task is set up. For example, does one ask for a dichotomous choice (the sentence is either good or not good), or does one ask for an indication of relative "goodness" on a Likert scale? There is little uniformity in the second language literature on this issue. Consider a single issue of the journal *Studies in Second Language Acquisition* [2001, 23(2)]. In it there are two articles whose authors used a Likert scale. Inagaki, in an article on the acquisition of motion verbs, used a 5-point Likert scale (-2 = completely unnatural, 0 = not sure, +2 = completely natural), whereas Montrul, in an article on agentive verbs, used a 7-point Likert scale (-3 = very unnatural, 0 = cannot decide, 3 = very natural).

Some researchers elect not to allow the "not sure" option or the middle of the road option and instead have a 4-point scale (see Hawkins & Chan, 1997). Juffs (2001) pointed out that without a standard in the field, it is difficult to compare the results of studies. He also noted that having a positive and negative scale with a zero midpoint makes it difficult to interpret a zero response as a "don't know" or as a midpoint. His suggestion was to use a completely positive scale (e.g., 1-7). Another possibility is to put no numerical value on the scale itself and to use a scale with only descriptors, as shown here:

X.....X.....X.....X.....X.....X.....X.....X

Very natural

Don't know

Very unnatural

When doing the actual scoring, one needs to separate the grammatical from the ungrammatical sentences in order to determine learners' knowledge of what is correct (i.e., grammatical) and knowledge of what is ungrammatical (i.e., excluded by their second language grammar). If one is using a 5-point Likert scale, the following is a possible scoring scheme:

<i>Grammatical Sentences</i>	<i>Ungrammatical Sentences</i>
Definitely correct = 4	Definitely incorrect = 4
Probably correct = 3	Probably incorrect = 3
Don't know = 2	Don't know = 2
Probably incorrect = 1	Probably correct = 1
Definitely incorrect = 0	Definitely correct = 0

One might also give partial credit for recognizing where the error is, even though the correction may not be accurate. Before doing the actual scoring, it is important to consider the corrections that have been made. For example, assume that you are studying the acquisition of morphological endings and the following sentence appears: *The man walk to the subway*. If the learner marks this "Incorrect" and then changes the second definite article on the test sheet to *a*, one would want to ignore the correction and count the sentence as if the learner had said "Correct" because the object of inquiry, morphological endings, was deemed to be correct.

Although acceptability judgments had their origins within formal linguistic approaches to language and are most commonly used within the UG paradigm, like other elicitation techniques they are not limited to that paradigm. Whatever the research question, they are difficult to use and interpret and must be employed with caution and careful thought as well as with awareness of the advantages and difficulties.

3.3.2. Elicited Imitation

Elicited imitation, like acceptability judgments, is used to determine grammaticality. The basic assumption underlying elicited imitation is that if a given sentence is part of one's grammar, it will be relatively easy to repeat. It is as if sentences are "filtered" through one's grammatical system. In elicited imitation tasks, sentences are presented to participants auditorily (i.e., either taped or orally by the researcher), and the participants are then asked to repeat them. The sentences are typically designed to manipulate certain grammatical structures, and a person's ability to repeat the sen-

tences accurately is a reflection of his or her internal grammatical system. A crucial factor in designing suitable test sentences is to keep the length at an appropriate level, generally one that exceeds short-term memory. Thus, sentences that might be appropriate for a more advanced level might be inappropriate for early-level learners. Elicited imitation, then, elicits an actual prompted utterance. This is unlike acceptability judgments, which elicit learners' beliefs about the language being learned. Following are recommendations for elicited imitation tasks:

1. Ensure an appropriate length in terms of words and syllables for all sentences. For example, a length between 12 and 17 syllables might be appropriate, depending on proficiency level.
2. Prerecord sentences for uniformity.
3. Randomize all sentences.
4. Include enough tokens of each grammatical structure so that you can make reasonable conclusions. This will depend on how many structures with which you are dealing. As with other methodologies, one has to balance the need to have an appropriate number of tokens with the necessity of not tiring the participants to the point that their responses are not reliable. Different randomizations for different learners can help guard against this latter possibility.
5. Ensure that there is enough time between the end of the prompt and the time that a learner begins to speak. (Sometimes researchers ask participants to count to 3 before beginning to speak to ensure that "echoic" memory is not being used.)
6. Pilot test everything.

3.3.3. Magnitude Estimation

Magnitude estimation is a well-established research tool used in a variety of disciplines (see Bard, Robertson, & Sorace, 1996, for a detailed description and history). It is useful when one wants not only to rank items in relation to one another, but also to know how much better X is than Y. It has been used when eliciting grammatical knowledge not as an absolute (yes or no), but as a matter of gradation (i.e., which sentence is more acceptable than another?). We can easily rank a list of things into an order of 1 through 9, but magnitude estimation allows us to determine whether each of the rankings is equidistant from the others, and, if not, the magnitude of the

ranking differences. Two of the positive aspects of this method, as noted by Bard et al. (p. 41) are:

- Researchers do not set the number of values that are used to measure the particular property of concern. Thus, the research does not impose a preset number of categories. The end result is a set of data that is more informative because the participant establishes both the range and the distribution of responses.
- One can observe meaningful differences that directly reflect differences in impressions of the property being investigated. This is so because magnitude estimation allows researchers to subtract the score on one sentence from that of another and be confident about the magnitude of difference.

As mentioned earlier, magnitude estimation is a ranking procedure. The scale that one uses is not imposed by the researcher, but rather is determined by each participant him- or herself. A stimulus is presented (orally or visually) and each participant assigns a numerical value. Each subsequent stimulus is rated according to the basis established from the previous stimulus. Thus, if a rater gives a value of 20 for an initial stimulus, and the second stimulus is perceived as being twice as good, he or she would give it a 40. It is common to train raters on the physical stimulus of line length. To do this, one shows raters a line and asks them to assign a numerical value to it. Following this, they are shown another line and are asked to assign a number to it in comparison to the length of the previous line. To make sure that raters understand the task, this can be repeated. It is best to tell raters to begin with a scale larger than 10 so that subsequent ratings do not end up with small numbers that might be difficult to accommodate. In other words, because ratings are multiples of a previous rating, if one were to start with 2, and the second were given half that value, the second one would have a value of 1. The next one, if smaller, would end up being a fraction, which, of course, could soon be an unwieldy number.

Following is a description of magnitude estimation in a second language study concerned with the effects of task repetition on language development (Gass, Mackey, Alvarez-Torres, & Fernández-García, 1999). The purpose of the judging was to rank excerpts of second language speech. The data were from English-speaking learners of Spanish who had repeated a narration of a film. The question was whether a later narration was better than an earlier one, and if so, by how much:

Description

After working with line length and before listening to the actual tapes, raters listened to a training tape. They heard three samples of 1 minute each which they rated using the magnitude estimation methodology. For the actual rating, raters listened to the first 2.5 minutes of each participant's tape. For purposes of analysis, to compare the magnitude of improvement judged by each rater it is necessary to convert the unique scales created by individual raters into a logarithmic scale. Conversion to a logarithmic scale is standard procedure when using magnitude estimation. Because the methodology allows for unique scales to be created by each rater, there must be a way to standardize the scales across raters to obtain a meaningful comparison. (Gass et al., 1999, p. 560)

As with all procedures, instructions are important. Following are the instructions provided to the raters of the magnitude estimation test in the previous study.

Instructions

You will hear nine tapes of different nonnative speakers of Spanish doing an on-line description in Spanish of a video they were watching. Your task is to rate their Spanish. Assign any number that seems appropriate to you to the first speech sample. This number will be your "base." Then assign successive numbers in such a way that they reflect your subjective impression (use a range wider than 10). For example, if a speech sample seems 20 times as good, assign a number 20 times as large as the first. If it seems one-fifth as good, assign a number one-fifth as large, and so forth. Use fractions, whole numbers, or decimals, but make each assignment proportional to how good you perceive the person's Spanish to be. (Gass et al., 1999, p. 581)

3.3.4. Truth-Value Judgments and Other Interpretation Tasks

Truth-value judgments are a way of understanding how people interpret sentences. These have been used extensively in the study of reflexives by L2 learners. An example of a truth-value token from Glew (1998) follows:

Bill was sick and in the hospital. Nobody knew what was wrong with Bill. The hospital did a lot of tests on Bill to find out what was wrong. Bill had to wait a long time in his hospital room. Finally, a doctor came in to tell Bill why he was sick.

After the medical tests, the doctor informed Bill about himself. (pp. 99–100)

True

False

Provided with sufficient contextual information, participants are able to consider all possible referents for *himself* (Bill or the doctor). In other words, appropriateness is determined by the context of the story.

Creating stories of this sort is a difficult process, and all such stories should be piloted. To underscore the importance of pilot studies, here is another example that was created for a study on reflexives (Glew, 1998), but ruled out after preliminary testing because there were multiple interpretations:

Sally drove Jane to a party. They had a good time, but Sally had too much to drink. Jane didn't want her to drive home so Jane offered to drive.

Sally was happy that Jane drove herself home.

True

False

This example was intended to elicit "False," because Jane drove Sally home (not herself), but it is clear that the story could be interpreted either way. Needless to say, this example was not included in the final set of materials.⁶

3.3.5. Sentence Matching

Sentence matching is a procedure that, like acceptability judgments, has its origins in another discipline, in this case psycholinguistics. Sentence matching tasks are usually performed on a computer. Participants are seated in front of a computer and are presented with a sentence that is either grammatical or ungrammatical. After a short delay, a second sentence appears on the screen, with the first sentence remaining in place.

⁶Over the years there have been other means of obtaining information about reflexives. This is a particularly difficult structure to investigate because many sentences are only grammatical given a particular context. Other than the truth-value paradigm discussed here, researchers have used multiple-choice formats.

Example: John said that Bill hit himself.

Who does *himself* refer to?

- a. John
- b. Bill
- c. Either John or Bill
- d. Another person
- e. Don't know

Lakshmanan and Teranishi (1994) pointed out that this is not an acceptable task because we gain information about whom *himself* can refer to, but not about whom *himself* cannot refer to. They offered the following revision (see original article for their interpretation of results, p. 195):

John said that Bill saw himself in the mirror.

- a. "Himself" cannot be John. agree disagree
- b. "Himself" cannot be Bill. agree disagree

For further discussion of many of the methodological points with regard to the study of reflexives, see Akiyama (2002).

Participants are asked to decide as quickly as possible if the sentences are identical or are not identical (i.e., if they match or do not match), entering their decision by pressing specific keys. The time from the appearance of the second sentence to the participant's pressing the key is recorded and forms the database for analysis. Research with native speakers has shown that participants in a matching task respond faster to matched grammatical sentences than they do to matched ungrammatical ones. (See Gass, 2001, for possible explanations for this phenomenon.) In other words, it would be expected that the reaction time would be less for the following two sentences:

John stated his plan to steal the car.

John stated his plan to steal the car.

than for the following:

John stated his plan for steal his car.

John stated his plan for steal his car.

With regard to general design, there are many issues that have to be decided when doing a sentence matching task, and researchers are not in agreement as to the "best solution." Following is a list of some of the variables that need to be weighed when designing a study using a sentence matching task:

- How long the two sentences remain on the screen.
- Delay time between the two sentences.
- Whether or not the two sentences remain on the screen until the participant has responded.
- Whether or not the screen goes blank after a predetermined time.
- Whether standard orthography or upper-case letters are used.
- The physical placement of the second sentence relative to the first.
- Whether participants are provided with feedback after each response.
- How the keys are labeled (same, different; different, same).
- The number of items included.
- The number of practice items included.
- Whether participants control the onset of each pair of sentences.

Another consideration when using sentence matching tasks relates to which data are kept in the final data pool. For example, a participant might just press the “same” key with no variation. Given that this individual’s data are probably not reflective of anything other than his or her not being on task, a researcher might decide to eliminate these data. We discuss these issues in greater detail in chapter 5 (for further information on scoring, see also Beck, 1998; Bley-Vroman & Masterson, 1989; Duffield, Prévost, & White, 1997; Eubank, 1993).

3.4. PROCESSING RESEARCH

Processing research has its basis in psycholinguistic processing rather than in the structure of linguistic forms. As discussed earlier in this chapter, research on formal models of language emphasizes constraints on grammar formation, whereas in psycholinguistics the emphasis is on the actual mechanisms involved in learning. Clearly, there is overlap in the interests of both areas, but each paradigm (formal models and processing) has its own particular approach.

3.4.1. Sentence Interpretation

One model dealing with sentence interpretation is known as the Competition Model (Bates & MacWhinney, 1982). The Competition Model has spurred a great deal of research that focuses on how learners process information. The major concern is what information people use in coming to an understanding of the relationships of words in a sentence. For example, when we see or read a sentence such as *Sally kissed John*, how do we come to an interpretation of who kissed whom? In English, we rely on word order (the first noun is typically the subject), meaning and animacy status of lexical items (if the sentence were *The pencil kissed John*, we would be confused as to the relationship), and morphological agreement. Some languages use case markings as the dominant cue with word order being less important. Not all languages use these same criteria (called *cues*), and not all languages assign the same degree of importance or strength to each criterion.

The methodology in the second language research literature based on the Competition Model utilizes sentences with various cues. Learners whose native language uses cues and cue strengths that differ from those of

the target language are presented with sentences designed to contain conflicting cues and are asked to determine what the subjects or agents of those sentences are. As with many elicitation methods, there is a great deal of variation in the procedures. Some issues to consider are:

- Are sentences read or tape-recorded? There is a lack of consistency when sentences are read as it is often difficult to neutralize natural intonational biases. This is particularly important when using unusual sentences such as *The pencil the cat saw*.
- How many sentences? As with all other methods, in order to eliminate fatigue and avoid compromising the reliability of the study, it is usually necessary to limit the number of sentences. A study by Sasaki (1997) used 144 sentences, although most studies have used between 27 and 54.
- What is the pause time between sentences? There appears to be no widely accepted standard.
- What are the instructions? A typical set of instructions was presented in Harrington (1987): “You are going to hear a tape with a series of very simple sentences. After each sentence is read you will have to interpret it: you should say which one of the two nouns in the sentence is *the subject of the sentence*, that is, *the one who does the action*” (p. 360). In the Harrington study, half of the participants were given the “syntactic bias” instruction first (the subject of the sentence), whereas the other half were given the “semantic bias” instruction first (the one who does the action).
- In what format are the responses given: oral or written?

All of these areas are important to consider when designing a study in which sentence interpretation is being used as an elicitation technique. The advantage of sentence interpretation is that researchers can learn what cues learners use in comprehending second language sentences and how those cues might be related to first language knowledge.

3.4.2. Reaction Time

Reaction time is considered here because it is believed to shed light on how people process certain parts of language. It is assumed that the more time it takes to respond to a sentence, the more processing “energy” is required. For example, if someone is asked about the acceptability of sentences in

English, it would be predicted that a sentence such as *I saw a big beautiful cat today* (7 words) would take less time to respond to than a sentence such as *Who did Ann say likes her friend?* (7 words), because the second sentence represents a more complex syntactic structure (and, hence, a greater processing load) than does the first. Reaction time measures are often used in conjunction with other kinds of research already discussed. For example, they can be an integral part of sentence matching experiments because the framework underlying sentence matching relies on comparing reaction times between grammatical matched sentences and ungrammatical matched sentences. Times are generally measured in milliseconds.⁷

3.4.3. Moving Window

The moving window technique is another elicitation measure that is typically carried out on a computer. Like other data collection methodologies in second language research, it also has its roots in the discipline of psycholinguistics (Just, Carpenter, & Wooley, 1982). In a moving window experiment, words are presented on a screen one at a time, with each successive word appearing after a participant indicates that she or he is ready. In other words, it is the participant who controls the reading speed. Once a

⁷In addition to self-made programs, there are commercially available programs for measuring reaction times as well as for doing psycholinguistic research in general. Here are two:

- *E-Prime* (Version 1.1) [Computer software]. Pittsburgh, PA: Psychology Software Tools, Inc. and PsyScope. E-Prime is a graphical experiment generator used primarily in psychology research. It includes applications for designing, generating, and running experiments, as well as collecting, editing, and analyzing data. Data can also be exported to external statistical tools such as SPSS. The graphical environment allows users to select and specify experimental functions visually, and the paradigm wizard provides basic experimental models that users can modify to fit their goals. The package includes E-Studio (the graphical interface), E-Basic (a customizable scripting language into which the graphical representations are compiled), E-Run (an application affording stimulus presentation and data collection precise to the millisecond), E-Merge (an application allowing the combination of single-session data files into multisession data files, and keeping a data file history), E-DataAid (a data management feature allowing users to filter, edit, and export their data), and E-Recovery (a backup mechanism in case data is unexpectedly lost or corrupted).
- *PsyScope*. (n.d.). Retrieved November 9, 2003, from <http://pyscope.psy.cmu.edu/>

Similar to E-Prime, PsyScope (designed for Macintosh computers) is a program for the design and control of experiments, primarily used in psychology research. As of 2003, PsyScope development has ceased although the software is compatible with all Mac systems from OS7 to OS9, including the Classic environment in OSX. The program is compatible with all Apple hardware produced in the last six years. PsyScope 1.2.5 is available free of charge (although unsupported).

new word appears on the screen, the previous word disappears. After the entire sentence has appeared, the participant presses a button to indicate whether the sentence is grammatical or ungrammatical. Juffs and Harrington (1995) used a moving window technique to investigate differences between long-distance object extraction (*Who did Jane say her friend likes?*) and subject extraction (*Who did Ann say likes her friend?*). Their main concern was to investigate the source of any differences, focusing on both processing time and linguistic knowledge (acceptability judgments were also used in their experiment).

Moving window techniques can provide information about processing times for various parts of the sentence. Instructions for these tools are similar to those exemplified earlier for acceptability judgments. An example follows from Juffs and Harrington (1995; see also the instructions from Bley-Vroman et al. in sec. 3.3.1.2.). This is an acceptability judgment task, but with a focus on processing time. Because the participant indicates readiness to move to each subsequent word, researchers can determine which parts of a sentence require additional processing time:

Instructions

Speakers of a language seem to develop a “feel” for what is a possible sentence, even when they have never been taught any particular rules. For example, in English, you might feel that sentences (a) and (c) sound like possible sentences, whereas (b) and (d) do not.

- a. Mary is likely to win the race.
- b. Mary is probable to win the race.
- c. It seems that John is late.
- d. John seems that he is late.

In this experiment, you will read sentences word by word on a computer screen. Concentrate on how you feel about these sentences. Native speakers of English often have different intuitions about such sentences, and there are no right or wrong answers. Tell us for each one whether you think it sounds possible or impossible in English.

Read each sentence carefully before you answer. Think of the sentences as spoken English and judge them accordingly. Work as quickly and accurately as you can. (Juffs & Harrington, 1995, p. 515)

Essential components of the instructions, as for acceptability judgments, include an explanation of what *intuition* means, together with the fact that there are no right or wrong answers.

3.5. INTERACTION-BASED RESEARCH

The previous two research areas (formal models and processing research) have methodologies that stem from other disciplines (for the most part formal linguistics and psycholinguistics, respectively). We now turn to interaction-based research, in which the focus is learners' conversational interactions with others (e.g., other learners, native speakers, and teachers) and the developmental benefits of such interactions.

Within interaction-based research, the goal is usually to manipulate the kinds of interactions in which learners are involved, the kind of feedback they receive during interaction, and the kind of output they produce in order to determine the relationship between the various components of interaction and second language learning. The most common way of gathering data is to involve learners in a range of carefully planned tasks.

There are a variety of ways of categorizing task types (see Pica, Kanagy, & Falodun, 1993, for task categorization suggestions). For example, a common distinction is to classify tasks as one-way and two-way. In a one-way task, the information flows from one person to the other, as when a learner describes a picture to his or her partner. In other words, the information that is being conveyed is held by one person. In a two-way task, there is an information exchange whereby both parties (or however many participants there are in a task) hold information that is vital to the resolution of the task. For example, in a story completion task, a learner may hold a portion of the information and must convey it to another person before the task can be successfully completed.

Another way to classify tasks is to consider the resolution of the task. Is there one correct outcome, as in a closed task (e.g., when two learners need to identify exactly five differences between two pictures), or do the participants need to agree on a common outcome or conclusion, as in an open task such as a discussion activity?

Considering these dimensions, researchers need to be creative in eliciting interactive data. Frequently, one is interested in eliciting certain grammatical structures, with the idea that interactive feedback on nontargetlike forms might be associated with learning, possibly reflected through changes in the learners' output on the particular structures about which they have received feedback. Thus, it is always important to pilot whatever instrument is selected to make sure that opportunities for the production of appropriate forms and feedback are being provided. Before turning to

general ways of eliciting interactive data, a word is necessary about recording data.

When recording data, the most important piece of equipment is a good microphone for a tape recorder. Clip-on microphones allow voices to be easily heard on the tape. Omnidirectional microphones may work if the research is being carried out in a laboratory (i.e., research with only the participants in a room); however, if the research is classroom based, it may be essential to use clip-ons. Ideally, tape recorders should have two inputs for microphones or, alternatively, an adapter can be used. Having two inputs makes later transcription easier. More detailed information about the recording and transcribing of oral data can be found in chapter 8. In the current section we describe commonly used data collection techniques within the interaction paradigm. In each case, for purposes of exposition, we categorized tasks. This is an unavoidable oversimplification; in many instances, there is overlap between or among the task types.

3.5.1. Picture Description Tasks

Many picture description tasks are information-gap tasks. Successful task completion usually depends on learners sharing information. In many such tasks it is important to ensure that if someone is describing a picture to another, the describer's picture cannot be seen. When this is the case, individuals (usually two, although these tasks can also be carried out in small groups) are separated by some barrier. This barrier can be made of cardboard or can even be a simple file folder. Whatever is used, the point is to ensure that none of the picture can be seen through the back. In some versions of picture description tasks, one person is given a picture with instructions to describe the picture so that another person can draw it. Instructions must also indicate that the person with the picture cannot show the picture to the other person. In some instances, such as when one wants to manipulate different types of input, recorded instructions and descriptions may be appropriate. There are experimental contexts in which one might want standardized input—that is, input that is the same for all participants—for instance, modified input (e.g., Gass & Varonis, 1994).⁸ In this situation, researchers can prepare a tape or a transcript and then use that tape or tran-

⁸Modified input is often associated with the way one might talk to someone who has limited abilities in the second or foreign language, using simpler vocabulary or speaking more slowly, for example. This is in contrast to unmodified input, which refers to input that would be spoken to a fluent speaker of the language.

script in the actual experiment. Other kinds of picture description tasks are collaborative and do not involve a gap in information. An example of this type of task is a dictogloss task, described in section 3.5.4.

3.5.2. Spot the Difference

Spot the difference tasks utilize pictures that are different in predetermined ways. Participants are asked to find the differences, and the number of differences can be prespecified so that the participants have a goal toward which to work. As with picture description tasks, it is important to ensure that the pictures are appropriate in terms of vocabulary for the level of the participants. In terms of format, it is crucial that participants not see their partners' pictures. An example of a setup is seen in Fig. 3.1 (this setup can also be used for picture description tasks, or any sorts of tasks where participants should not view each other's pictures).

Figures 3.2 and 3.3 show examples of pictures that can be used for spot the difference tasks between two or three participants. Figure 3.2 reflects a



FIG. 3.1. A simple barrier placed between two participants.

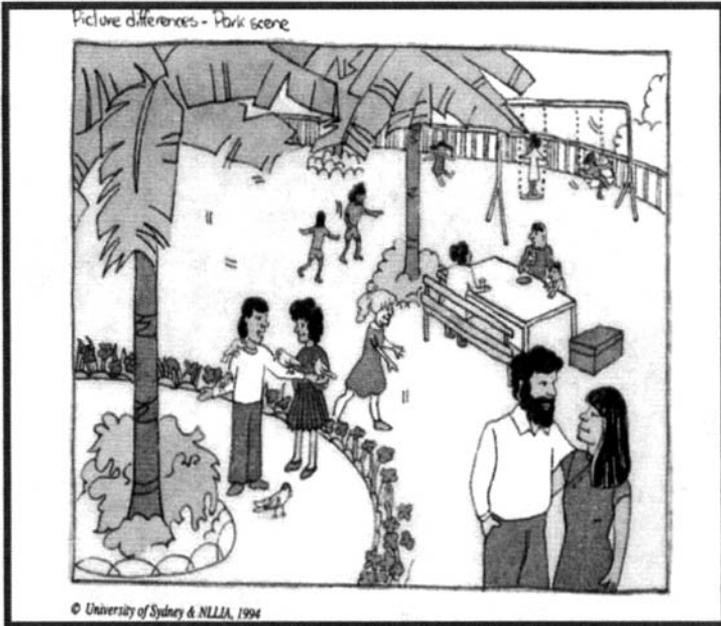


FIG. 3.2. Spot the differences park scene. *Communicative Tasks* (1995). National Languages and Literacy Institute of Australia, Language Acquisition Research Centre, University of Sydney. © National Languages and Literacy Institute of Australia (NLLIA). Reprinted with permission.

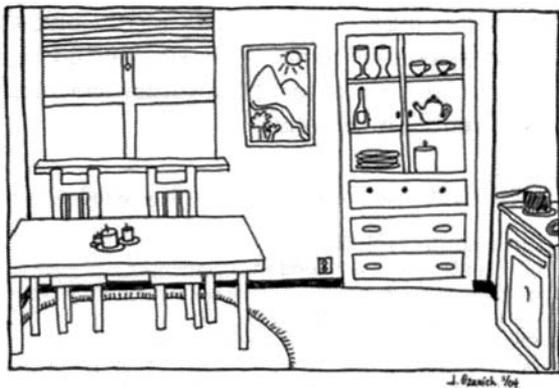
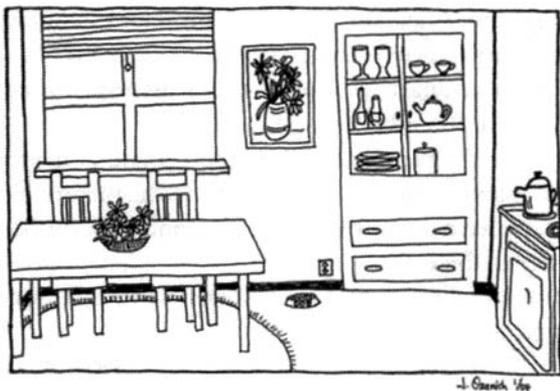
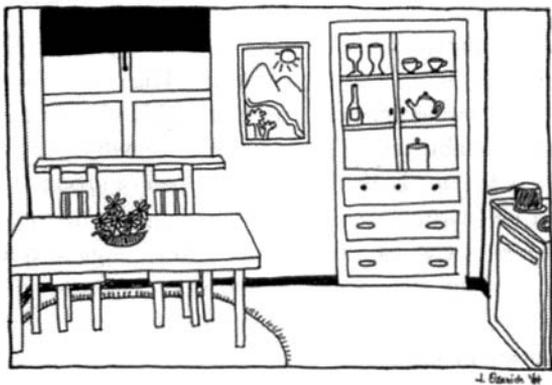


FIG. 3.3. Spot the differences kitchen task.

park scene. These pictures can be used to elicit locatives, plurals, and, as with most spot the difference tasks, questions. The vocabulary is somewhat difficult (e.g., *slide*, *swing*), but this can work with advanced learners or by using pretaught vocabulary. The picture is somewhat “busy” but can be modified to meet the needs of an appropriate pair or group of participants.

Figure 3.3 shows a kitchen scene that can be used with three participants (or two, if one uses two of the three pictures).⁹ Again, question forms, locative constructions, plural forms, and (easier) household vocabulary would be produced. Following are guidelines for spot the difference and picture description tasks:

1. Find a picture that contains items that can easily be described, but that includes vocabulary that is likely to cause some lack of understanding, and hence, some negotiation. This might involve physical objects or the placement of objects (*above*, *on top of*).
2. Separate individuals by a barrier or at least ensure that the picture is not visible to the other person in the pair.
3. Ensure that the picture contains appropriate items for description and/or appropriate location of items in the picture. For example, a picture with a car on top of a house would add another element of difficulty to the task.
4. If relevant, make sure that the task elicits the linguistic structures or forms of interest.
5. Ensure that there are sufficient opportunities for interactional modifications, feedback, and output based on the research question.
6. For picture description, make sure the participants understand that the person drawing should not see the picture until the task is completed. For spot the difference, make sure that no participant shows his or her picture to the other(s). Inform participants about the number of differences, if necessary.

⁹Below is a list of the differences that participants can be asked to identify in the kitchen task (Fig. 3.3):

1. Shades (2 are striped/1 is black).
2. Picture (2 are landscapes/1 is a floral arrangement).
3. Table centerpiece (2 are flowers/1 is candles).
4. Electrical outlet (2 have one/1 does not).
5. Bottle in cabinet (1 has two on 2nd shelf/2 have only 1).
6. Plates in cabinet (2 have plates on bottom shelf/1 has no plates).
7. Drawers in cabinet (2 have three drawers/1 has two drawers).
8. Oven window (2 have windows/1 does not).
9. Pot on stove (2 have a pot on the stove/1 has a tea kettle).
10. Dog food dish (1 has one on the floor/2 have no dog dish).

7. As usual, carefully pilot test the task.

3.5.3. Jigsaw Tasks

In a jigsaw task, which is a two-way task, individuals have different pieces of information. In order to solve the task, they must orally interact to put the pieces together. One example of a jigsaw task is a map task¹⁰ in which participants are given a map of a section of a city. Each participant is provided with different information about street closings, and they must explain to each other about which streets are closed and when. Once this portion is completed, they have to work together to determine a route from Point A to Point B by car, keeping in mind that certain streets are closed. The example in Fig. 3.4 below shows a map of a city with locations in Spanish. Examples of sets of street closings are also included, along with an English translation. Alternatively, each person can be given a map with pre-

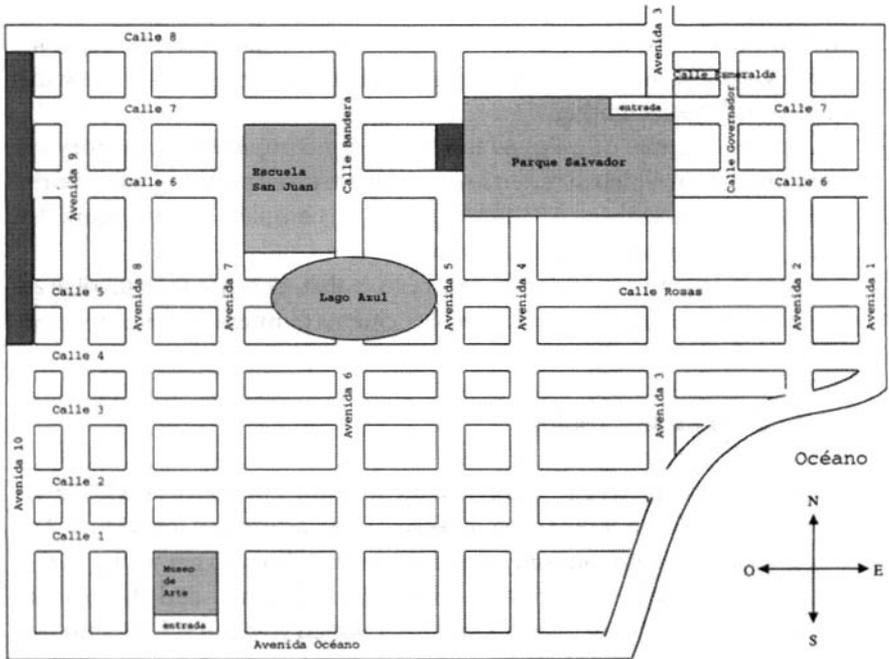


FIG. 3.4. Map task.

¹⁰For other examples of jigsaw tasks, see Pica et al. (1993).

Example of Instructions—Map Task

<i>Participante 1</i>	<i>Participante 2</i>
<ul style="list-style-type: none"> ■ La avenida 10 está cerrada entre la calle 4 y la calle 8. <i>Avenue 10 is closed between Street 4 and Street 8.</i> 	<ul style="list-style-type: none"> ■ La avenida 5 está cerrada entre la calle 6 y la calle 7. <i>Avenue 5 is closed between Street 6 and Street 7.</i>
<ul style="list-style-type: none"> ■ La calle 5 está cerrada desde el Lago Azul hasta la avenida 10. <i>Street 5 is closed from the Blue Lake to Avenue 10.</i> 	<ul style="list-style-type: none"> ■ La avenida 8 va en una sola dirección hacia el sur. <i>Avenue 8 goes in a single direction toward the south.</i>
<ul style="list-style-type: none"> ■ La avenida Oceano va en una sola dirección hacia el oeste. <i>Ocean Avenue goes in a single direction toward the west.</i> 	<ul style="list-style-type: none"> ■ La avenida 2 estará cerrada todo el día. <i>Avenue 2 will be closed all day.</i>

blocked-off streets. In this instance, participants would receive a separate blank map in order to draw the route, with instructions not to show the original map to each other.

Another example of a jigsaw task is a story completion, or a story sequencing task, in which different individuals are given parts of a story (written or pictorial) with instructions to make a complete story. Figure 3.5 provides an example of this sort of task.¹¹

The important point about jigsaw tasks is that, because they involve an information exchange, they require participants to interact while completing the task.

3.5.4. Consensus Tasks

Consensus tasks generally involve pairs or groups of learners who must come to an agreement on a certain issue. For example, ten individuals are stranded on an island, but only five can fit into a boat to get to the mainland. Characteristics are provided for each individual, and the pair or group must come to an agreement about which five should get into the

¹¹ Answer key to the story sequencing task (Fig. 3.5): A robber stole a wallet while two students were away playing tennis. The robber was chased by the dog and dropped the wallet. The student found the wallet while gardening a year later—he didn't know that the dog had buried it.

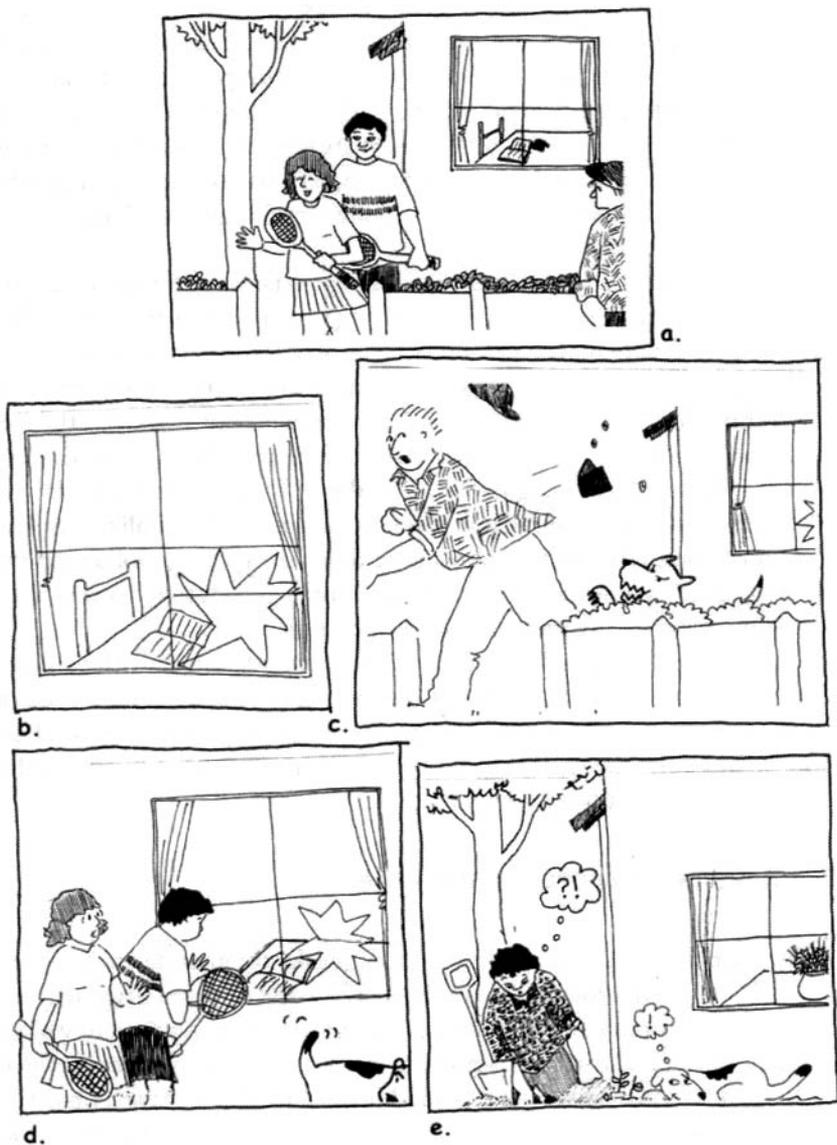


FIG. 3.5. Story sequencing task.

boat. This task allows for a less guided discussion than do other tasks, but it does not guarantee that there will be interaction. One individual might not participate, or, if the task is not engaging, participants might take only a few minutes to pick five individuals without giving elaborate justification. As with other methods, instructions are important to ensure that the participants understand the need to participate. For example, each participant can be assigned a role of an individual and argue for that person's suitability for the boat.

Another type of consensus task is a dictogloss task (see Swain & Lapkin, 2001). In this type of task, learners work together to reconstruct a text that has been read to them. It is possible to choose a text based on content, vocabulary, or particular grammatical structures. In its normal mode of delivery (although this could be modified for the purposes of research), the text is read aloud twice at normal speed. Participants can take notes on the first reading, the second reading, both, or neither. This will depend on the researcher's goals. Because the text is read at normal speed (unlike a typical dictation), the participants cannot write down everything. Following the readings, participants can work in dyads or small groups to reconstruct the text while maintaining the meaning of the original.

One concern about task-based research is that different activities and different interactive treatments might involve more or less time on task, or more or less input and output of linguistic form and interactional modification. It is useful to keep this in mind when planning research. Again, pilot testing is crucial in this respect.

3.5.5. Consciousness-Raising Tasks

As the name implies, consciousness-raising tasks are intended to facilitate learners' cognitive processes in terms of awareness of some language area or linguistic structure. In these tasks, learners are often required to verbalize their thoughts about language on their way to a solution. An example of such a task was provided by Fotos and Ellis (1991). The linguistic object of study was dative alternation (*I gave books to my friends* versus *I gave my friends books* and *I suggested a plan for her* but **I suggested her a plan*). Each student in a group of four was given a sentence (some were correct and some not). Their task was to talk and determine the "rules" for dative alternation in English. Each student read his or her sentence aloud to the others and then, using a worksheet, the students determined which verbs could and could not use the alternating structure.

3.5.6. Computer-Mediated Research

Computer-mediated communication (CMC) involves learners in communicative exchanges using the computer. CMC is a text-based medium that may amplify opportunities for students to pay attention to linguistic form as well as providing a less stressful environment for second language practice and production. Thus, it may be that CMC can provide richer data for second language learners than can face-to-face oral exchanges. CMC software generally allows users to engage in both simultaneous (chat-based) and asynchronous (forum-based) communication. What is typed is stored, and users, teachers, and researchers can then retrieve prior conversations if desired. The forums generally include open, moderated, closed, and restricted formats, and some have support for distance learning.

Computer-based research can also utilize the various tracking possibilities that technology allows—for example, to see the extent to which learners do and do not use look-up sources, and when they do, how and how often. This can then be examined in the context of measures of learning. Because computer-based data allow learners more anonymity than do face-to-face data, they may also be less restricted in what they say and how they say it. Again, this can be examined in the context of learning.

There are various computer programs that can be used in second language research. As one attempts research using computers, it is important to be informed about advances in technology that will best match one's research question.

3.6. STRATEGIES AND COGNITIVE PROCESSES

Strategy-based research is aimed at determining the strategies used when learning a second language together with the variables that determine the selection of strategies. Macaro (2001, p. 37) listed the following ways of gaining access to that information:

- Asking learners (directly or through a questionnaire) which strategies they use in general, or which strategies they use when attempting a particular task.
- Observing learners while they work at their language learning tasks.
- Asking learners to give a *retrospective commentary* on how they learn (e.g., through keeping diaries or through dialogue journals).

- Asking learners to provide a *synchronic commentary* on how they accomplished a task (i.e., to talk about their thoughts to an interviewer while they were doing the task).
- Tracking learners (through the use of a computer) on a variety of measures.

In the following sections, we elaborate on some of the resources that researchers employ to elicit strategy information, including observations and introspective methods such as think-alouds and stimulated and immediate recalls.

3.6.1. Observations

Observations, discussed in greater detail in chapter 7, frequently take place within a classroom context. Macaro (2001) provided ways we may consider conducting research on strategies used within the classroom, although there are clearly limits to the conclusions that can be arrived at on the basis of observational data. In terms of strategies-based research, we might observe:

- when students are moving their lips, which might be an indication that they are preparing themselves to speak by practicing under their breath,
- to what extent and which students are “buying processing time” by using such markers as ‘uh’ or ‘well’ or other discourse markers designed to show that they wish to keep their turn,
- to what extent students are employing the compensation strategy of circumlocution (finding alternative ways of saying something they don’t know how to say),
- which students are asking a friend for help when they don’t understand,
- which students are sounding out words before saying them,
- which students are reasoning by deduction (“it must mean this because of this”),
- which students are focusing on every word rather than the gist, perhaps by observing them as they move their finger from word to word,
- which students plunge straight into [an] activity and which students spend some time planning their work, and
- which students use the dictionary and with what frequency. (Macaro, 2001, p. 66)

3.6.2. Introspective Measures

Like other methods discussed thus far, introspective methods, which tap participants' reflections on mental processes, originated in the fields of philosophy and psychology. (See Gass & Mackey, 2000, for more information about the origin and use of introspective methods in second language research.) The use of introspection assumes that what takes place in consciousness can be observed in much the same way that one can observe events in the external world.

Verbal reporting is a special type of introspection and consists of gathering protocols, or reports, by asking individuals to say what is going through their minds as they are solving a problem or completing a task. Cohen (1998) outlined three primary types of verbal reporting used in second language research:

1. *Self-report*: With self-report data, one can gain information about general approaches to something. For example, a statement such as "I am a systematic learner when it comes to learning a second language" might be found on a typical second language learning questionnaire. Such statements are removed from the event in question and are less of concern here than are other types of verbal reporting.
2. *Self-observation*: Self-observation data can be introspective (within a short period of the event) or retrospective. In self-observation, a learner reports on what she or he has done. An example from Cohen (1998) is the following: "What I just did was to skim through the incoming oral text as I listened, picking out key words and phrases" (p. 34). Such self-observations refer to specific events and are not as generalized as self-report data.
3. *Self-revelation* (also known as "think-aloud"): A participant provides an ongoing report of his or her thought processes while performing some task (see sec. 3.6.2.2 for further details).

In general, we can think of introspective reports as differing along a number of dimensions: currency (time frame), form (oral or written), task type (think-aloud, talk-aloud, or retrospective), and amount of support for the task.

The major advantage to the use of verbal reports is that one can often gain access to processes that are unavailable by other means. However, it is also possible to question the extent to which verbal report data are valid and reliable.

The major disadvantage to the use of verbal reports as data has to do with the accuracy of the reporting. This is particularly the case in self-report and self-observational data. With such data, when the time between the event being reported and the reporting itself is short, there is a greater likelihood that the reporting will be accurate. This particular issue is discussed in relation to stimulated recall next. We further discuss introspective measures in chapter 7 on second language classrooms.

3.6.2.1. Stimulated Recall

Stimulated recall is usually viewed as a subset of introspective measures. It is a means by which a researcher, in an effort to explore a learner's thought processes or strategies, can prompt the learner to recall and report thoughts that she or he had while performing a task or participating in an event. Gass and Mackey (2000) provided an extensive description of stimulated recall, together with examples of its use (see Mackey, Gass, & McDonough, 2000, for an example of stimulated recall in an experimental context). Stimulated recalls are conducted with some degree of support; for example, learners may be shown a videotape so that they can watch themselves carrying out the task, or they may be given their second language written product so that they can follow the changes they made, commenting on their motivations and thought processes along the way.

One thing that is clear from the proliferation of studies using stimulated recalls—and from the corresponding number of critiques of verbal report methodologies—is that stimulated recall is a methodology that, like many others, must be used with care. Numerous potential problems relate to issues of memory and retrieval, timing, and instructions. Thus, studies that utilize stimulated recall methodology require carefully structured research designs to avoid pitfalls. Recommendations for stimulated recall research (adapted from Gass & Mackey, 2000) include:

1. Data should be collected as soon as possible after the event that is the focus of the recall. This is to increase the likelihood that the data structures being accessed are from short-term memory. Retrieval from long-term memory may result in recall interference, and as the event becomes more distant in memory, there is a greater chance that participants will say what they think the researcher wants them to say because the event is not sharply focused in their memories.

2. The stimulus should be as strong as possible to activate memory structures. For example, in a stimulated recall of oral interaction, participants can watch a video if the recall is immediately after the event. If it is more delayed, they can watch a video and possibly even read a transcript of the relevant episodes as well.
3. The participants should be minimally trained; that is, they should be able to carry out the procedure, but should not be cued into any aspects that are extra or unnecessary knowledge. This can be achieved through the use of pilots. Often, simple instructions and a direct model will be enough in a stimulated recall procedure. Sometimes, even instructions are not necessary; the collection instrumentation will be sufficient (e.g., in the case of a questionnaire or a Q-A interview).
4. How much structure is involved in the recall procedure is strongly related to the research question. Generally, if participants are not led or focused, their recalls will be less susceptible to researcher interference. Also, if learners participate in the selection and control of stimulus episodes and are allowed to initiate recalls themselves, there will again be less likelihood of researcher interference in the data. However, unstructured situations do not always result in useful data.

3.6.2.2. Think-Alouds or Online Tasks

In think-aloud tasks, also known as online tasks, individuals are asked what is going through their minds as they are solving a problem or completing a task. Through this procedure, a researcher can gather information about the way people approach a problem-solving activity. The following protocols, from the solving of a mathematical problem, illustrate two very different thought processes during the solving of a problem (from van Someren, Barnard, & Sandberg, 1994, pp. 5–6). A comparison of these two protocols reveals the way that two individuals can solve the same problem correctly, but in two vastly different ways. By looking only at the starting point and the end product (the solution), it would be difficult to fully understand the two different approaches that can be observed by comparing the complete protocols:

Problem to be solved: A father, a mother and their son are 80 years old together. The father is twice as old as the son. The mother has the same age as the father. How old is the son?

Student 1

1. a father, a mother and their son are together 80 years old
2. the father is twice as old as the son
3. the mother is as old as the father
4. how old is the son?
5. well, that sounds complicated
6. let's have a look

7. I just call them F, M and S
8. $F + M + S = 80$
9. $F = 2 \times S$
10. and $M = S$
11. what do we have now?
12. three equations and three unknowns
11. what do we have now?
12. three equations and three unknowns
13. so $S = \dots$
14. $2 \times S + S + S = 80$
15. so $4 \times S = 80$
16. so $S = 20$
17. $S = 20$
18. yes, that is possible

19. so father and mother are 80 minus 16
20. 64

21. er ... 32

Student 2

1. father, mother and son are together 80 years old
2. how is that possible?
3. if such a father is 30 and mother too
4. then the son is 20
5. no, that is not possible
6. if you are 30, you cannot have a son of 20

7. so they should be older
8. about 35, more or less
9. let's have a look
10. the father is twice as old as the son
11. so if he is 35 and the son 17
12. no, that is not possible
11. so if he is 35 and the son 17
12. no, that is not possible
13. 36 and 18
14. then the mother is
15. 36 plus 18 is 54
16. 26 ...
17. well, it might be possible
18. no, then she should have had a child when she was 9

19. oh, no
20. no the father should, the mother should be older
21. for example 30
22. but then I will not have 80
23. 80 minus 30, 50
24. then the father should be nearly 35 and the son nearly 18
25. something like that
26. let's have a look, where am I?
27. the father is twice ...

28. the mother is as old as the father
 29. oh dear
 30. my mother, well not my mother
 31. but my mother was 30 and my father nearly 35
 32. that is not possible
 33. if I make them both 33
 34. then I have together 66
 35. then there is for the son ... 24
 36. no, that is impossible
 37. I don't understand it anymore
 38. 66, 80
 39. no, wait, the son is 14
 40. almost, the parents are too old
 41. 32, 32, 64, 16, yes
 42. the son is 16 and the parents 32, together 80
-

Even though both participants arrive at the correct answer, different problem-solving approaches are revealed: one algebraic and one “hit or miss” combined with logic (how old a mother or father was likely to have been when the child was born). Considering only the outcome reveals nothing of the complexities involved in the means of getting there.

In second language research, an example of a think-aloud task can be seen in research by Leow (1998), who investigated issues of attention in second language learning. Leow used crossword puzzles as the task that learners had to perform. Following are the instructions he gave to his university-level learners of Spanish:

Instructions for Think-Alouds

Here is a crossword puzzle similar to the ones you have done in class ... Use the clues provided and see if you can successfully complete this crossword. As you do the crossword, try to speak aloud into the microphone your thoughts WHILE you perform the task for each word, NOT AFTER. Include the numbers of the clues also while you are thinking aloud. Please try to speak in a clear voice. (Leow, 1998, p. 137)

The kind of data that can be elicited through this method can be quite rich, as can be seen in the following excerpt from Leow (1998). The boldface print indicates words in Spanish:

Vertical now ... 2 down, OK I have an *o* here but I don't know why because in 1 across I have *se morio* but I guess it has to be **murio** because 2 down has to be *un* [changes *o* to *u*] ... OK I have to but it must be *tu* so it means that 7 across for the past tense of *dormirse* must be **durmio** instead of *dormio* [changes *o* to *u*] ... OK third person plural form of the verb *pedir* they asked for, 5 down ... *pedieron* [pause] OK I am wondering whether because I have **pidieron** [spells out] and I am thinking it should be *pe-* but that would make it *dormeo* with an *e* instead of *I*, ... I guess I will see how the other ones go and take a look at that one again ... OK, the opposite of *no* is *si* which means that for 11 across I have *mentieron* but it should be **mintieron** for the third person plural past tense of *mentir*, *mintieron* [changes *e* to *I*] which makes me now realize that, **pidieron** with an **I** is probably right since the *e* in *mentir* changes to an **I** so the *e* in *pedir* is also going to change to an **I** as well ... OK 12 down, the opposite of *no* is *s* (which means that where I have *corregio* it becomes **corrigo corrigo** so the third person singular of past tense *corregi* is **corrigo** [changes *e* to an *I*] ... looks like all the *e*'s are becoming *i*'s in the stems ... OK, third person singular form of *descubrir* discovered OK it is *descubrio*, OK 17 down possessive adjective in Spanish OK now here yet again I have *to* because I have *se dormieron* and that must become *tu* so it becomes *se **durmieron*** [changes *o* to *u*] OK third person singular form of *preferir* preferred, OK now here yet again *prefe-* [spells out] is going to change to **prefi-** [spells out] **prefirio** [changes *e* to *i*] ... OK 25 down, the verb to go in Spanish which is *ir* and I have *er* [spells out] because with 24 across I have *repetieron* but I guess now that *e* becomes an *i* becomes **repitieron** ... [changes *e* to *i*] ... and 25 down is *ir*, so now I am going to go back and change any other ones where I have *e* in the stem that should become an **i**, like 1 down, I believe would become *se **divirtieron***, it becomes an *I* and everything else looks OK so I guess that's it. [9 Minutes]¹²

The example from Leow shows how an individual thinks about a grammar problem. The following example from Morrison (1996) demonstrates a different use of think-alouds, this time focusing on talk between two individuals. Morrison's study was concerned with inferencing strategies used by L2 French

¹²From Leow, R. (1998). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) fine-grained analysis of attention. *Applied Psycholinguistics*, 19, 146. Copyright © 1998 by Cambridge University Press. Reprinted with the permission of Cambridge University Press.

learners when encountering unknown vocabulary words in an authentic reading passage. Learner A and Learner B read the text individually, and were asked to think aloud together about the meanings of the underlined words. The following excerpt is based on their discussion of *piétons*, “pedestrians”:

- A2 OK, “*piétons*.”
- B3 I think that’s *pedestrian*.
- A4 OK ...
- B5 I had no idea when I was writing ...
- A6 Me either, but I have to admit, unfortunately *pedestrian* didn’t—that’s good. I didn’t think of that—it’s so obvious. But for some reason I thought it had to do with something like ... just because it said modern urban city, so I thought of just like, sort of like *businessman*.
- B7 Mmm. (oh)
- A8 I thought there was some sort of French word for like modern city dweller.
- B9 Right.
- A10 I didn’t think of *pedestrian*. But that’s right. I think it’s *pedestrian*.
- B11 I think even, because you see ‘*pied*’ and you think foot.
- A12 I didn’t think of that. But, definitely.
- B13 And so, I think it has something to with that. But, again, in there, I had absolutely no idea. I think it’s even, when you read the rest of this, this ...
- A14 thing, that’s when you understand. That’s why now it makes sense.
- B15 Exactly
- A16 But I should have figured it out.
- B17 You ... different perspectives, you know think, because, you know they talk about being on the road on the street, and the way, you know, they carry themselves on the street and stuff, and right away you think, you know, *pedestrians*. Right?
- A18 Yeah.
- B19 Like, that’s who you’d find on the street.
- A20 Uh-huh. Yeah
- B21 So
- A22 Yeah, I think you’re right
- B23 I think it has something to do with that. So “*piétons*” (writing) *pedestrian*.
- A24 And I think also, you’re right, the “*pied*” and then even just “*pied*” *pedestrian*, for the ...
- B25 Yeah, like “*pied*,” “*pied*,” that’s what—

- A26 Yeah, OK
B27 ... so *pedestrian* ...
A28 And I guess also because you don't have *pedestrians* in the country, countryside. Like you wouldn't call it a pedestrian ... as in roads, as in ...
B29 Yeah, when you think *pedestrians*, you think ...
A30 ... modern cities
B31 the city, city life ...

This example illustrates the process of learners reporting their earlier thoughts, together with the new inferencing strategies they use to refine their understanding. These include using their L2 French knowledge (B11, A24) as well as various contextual clues from further on in the passage (A6, B13, B17) and their real-world knowledge (B19, A28). The example shows the development of their comprehension of this word through the integration of these strategies.

The following basic recommendations for think-aloud protocols are adapted from Macaro (2001):

- Give participants a specific task to perform (reading and writing tasks work best).
- Make sure that they understand what they have to do and that they are comfortable with the task. In general, they should be told that you want to know what they are thinking when they are performing the task.
- Find a similar task and demonstrate how a think-aloud works. An alternative is to ask them to practice with a different task. This is often preferable because if the researcher models the task, it is possible that the learner will use the particular strategy that the researcher has used.
- Have the tape recorder ready and start it.
- Students may need to be encouraged when there is insufficient talk-aloud data. Avoid using phrases like "Are you sure?" and "That's good." Instead, use only phrases like "What makes you say that?", "What made you do that?" (if, for example, they are looking up a word in a dictionary), "What are you thinking at this moment?", and "Please keep talking."
- Listen to the recording of the think-aloud process (after the session) and make a list of all the strategies used by the student, employing a predeveloped coding system.

As we noted earlier, in all research that relies on participants' giving information on their thought processes (whether stimulated recall or verbal

think-alouds), one needs to be aware that participants may not be aware of their processes and/or they may not wish to reveal them. Gass and Mackey (2000, chap. 4) provided more detailed information on the dos and don'ts of recalls, particularly stimulated recalls.

3.6.2.3. Immediate Recalls

Immediate recall is a technique used to elicit data immediately after the completion of the event to be recalled. It can be distinguished from stimulated recall in that it must occur immediately after the event (whereas stimulated recall may or may not occur immediately following the event and it does not involve a stimulus to talk from—e.g., videotape, audiotape, written product), and it can be distinguished from online recall in that it does not occur simultaneously with the event. For example, in an experiment involving interaction data, immediate recall can take place after one conversational turn (e.g., 10 to 15 seconds in length) during a conversational session. A stimulated recall would take place after the entire conversation, using a tape of the conversation as stimulus.

Online recall is difficult to implement in interaction research, but immediate recalls have been used by Philp (2003) and Egi (2003) in explorations of what learners notice about conversational feedback. In Philp's study, learners were instructed to verbalize thoughts they had during a conversational turn immediately after a recall prompt, which consisted of two knocking sounds. The knocking sounds occurred in three contexts: immediately after recasts of nontargetlike production of the linguistic items targeted in her study, after other errors, and after correct responses. As with all recalls, immediate recall can be conducted in the learners' L1, to allow them to fully express their thoughts, or in the L2. Either way, training in immediate recall is often essential to help learners get used to the technique. Immediate recalls may suffer from fewer of the problems of memory decay that can be a problem with stimulated recalls, yet immediate recall is arguably a more artificial task and may also interfere with subsequent task performance. As with all techniques, one must pilot the technique to ensure not only that it works with the particular group of learners that will be used, but also that it elicits the type of data needed.

3.7. SOCIOLINGUISTIC/PRAGMATICS-BASED RESEARCH

Both sociolinguistics and pragmatics are the study of language in context. Thus, they emphasize social and contextual variables as they affect the

learning and production of a second language. The underlying assumption is that second language data do not represent a static phenomenon; rather, second language production is affected by such external variables as the specific task required of a learner, the social status of the interlocutor, and gender differences, among others. The resultant claim is that learners may produce different forms that are dependent on external variables. Pragmatically based second language research deals with both the acquisition and use of second language pragmatic knowledge.

Sociolinguistic and pragmatics-based research studies (see Kasper & Rose, 2002) are difficult to conduct using forced-elicitation devices, given that both consider language in context, and yet, like many other areas, it is often necessary to require examples if one is to collect sufficient data to draw conclusions. If, for example, one wanted to gather data on rudeness, either in terms of production or interpretation, it might be difficult to collect enough tokens from which one could draw reasonable generalizations. Researchers must therefore create contexts that require the necessary tokens. There are certain commonly used methods for doing this, and we discuss them in the following sections.

3.7.1. Naturalistic Settings

Researchers can, of course, attempt to set up situations in which certain language events will recur. Two examples come from research on advising sessions and one from a teaching context. Fiksdal (1990) investigated high-stakes interviews (in which the issue was the visa status of international students in the United States), analyzing university-based immigration counseling to both native speakers and nonnative speakers. She was able to obtain the cooperation of the immigration counselor and the students to videotape the sessions. The context was constant (the counselor's office), and thus the author was able to make comparisons between the language used by the advisor to the native speakers and to the nonnative speakers and comparisons of the responses by native speakers and nonnative speakers. Another example is a study by Bardovi-Harlig and Hartford (1996) on advising sessions for graduate students in an applied linguistics graduate program. The sessions were audiotaped, and again, because the context was constant, comparisons could be made between native and nonnative speakers on such areas as suggestions or disagreements. Tyler (1992) also collected data for a qualitative discourse analysis of the spoken English discourse of a nonnative English speaking international teaching assistant,

comparing it with the discourse of a teaching assistant who was a native English speaker. She asked native speakers to judge comprehensibility of the discourse based on hearing transcriptions read by a native speaker. This methodology allowed for comparability of the discourse.

3.7.2. Elicited Narratives

For a variety of different purposes, second language researchers can benefit from eliciting narratives from learners. There are a number of ways to do this. One can, of course, ask participants to tell a story about some past event or future plans in order to elicit these tenses. For example, one could ask:

- Tell me what you did yesterday.
- Tell me about a typical day.
- Tell me how you like to spend your free time.
- Tell me about the town where you live.
- Tell me your plans for the summer vacation.

One problem with this approach is that despite the prompt, learners may opt for a different form. The following example comes from research designed to elicit past tense forms in Spanish (English L1). Participants viewed a picture for 1 minute, then turned the picture over and read the instructions as a prompt for the narration of events. Here is the Spanish prompt with the English translation:

Prompt in Spanish: Anoche el señor Gonzalez estaba leyendo un libro en su casa. Había una lámpara de lectura detrás de él. Aún tenía puesto el traje porque acababa de llegar del trabajo, pero no tenía puesto los zapatos. [Last night Mr. Gonzalez was reading a book in his house. There was a reading lamp behind him. He was still wearing his suit because he had just returned from work, but he had taken off his shoes.]

Following is a response from one participant; this response was typical in that despite the efforts of the researcher, past tense was not used:.

Bueno, mientras el señor Gonzalez está leyendo se ve [eh] que en el otro lado de la pared va caminando su esposa, pues se supone que es su esposa con un regalo muy grande, una caja muy grande que obviamente es un regalo ... [the story continues]. [Good, while Mr. Gonzalez is read-

ing, you see [eh] that on the other side of the wall his wife is walking, well you assume that it is his wife with a very big gift, a very big box that is obviously a gift]

Another point to consider when eliciting narratives is planning time.¹³ Does one elicit a narrative immediately after providing the learner with a stimulus? Or does one allow the learner time to think about what she or he will say? Planning can impact the quantity and quality of what is produced. What follows are some dimensions that researchers need to consider when eliciting narratives:

- Should learners have time to plan?
- If so, how much time?
- If planning time is allowed, should learners be allowed to make notes for themselves?
- If so, can these notes be used during the retelling?
- If relevant, how can the use of a particular linguistic form be elicited?

In the next sections, we briefly describe other common ways of eliciting stories and extended stretches of speech (see also chap. 6 on data gathering).

3.7.2.1. Silent Film

One way to elicit data is through the retelling of a silent film. The idea is to give learners a uniform prompt from which to speak. Usually, these film clips are relatively short (about 2 to 4 minutes) and allow the researcher to keep all information constant. There are a few areas of which to be wary:

- Films must be as culturally neutral as possible if one is to use them for learners of different languages.
- Films cannot be too long or too short. If they are too short, there will not be a sufficient quantity of data. If they are too long, the learner might get embroiled in the recall of events.
- Learners may need to believe that they are telling it to a person who has never seen the film.

¹³Clearly, the issue of planning is relevant to many production tasks and has been the focus of much recent work in the second language field (see Ellis, 2003, for a review).

There are also variations to consider in the administration of a task like this:

- Do participants tell the story as the film is playing, or do they tell it after the entire film has been shown?
- Do they write their response?
- Do they tell the story orally?
- If they tell the story orally, do they tell it to a tape recorder or to someone who they believe has not seen the film?

3.7.2.2. Film Strips With Minimal Sound

There are some films that have minimal dialogue (for example, see Gass et al., 1999; Skehan & Foster, 1999). These can be used in the same way as silent films by turning the sound off. It is important that learners not be influenced by the speech of either their native language or the target language. On the other hand, it is important that no essential dialogue be removed that would prevent participants from fully understanding the story.

3.7.2.3. Picture Tasks

In section 3.5.4, we discussed a type of consensus task that involved telling a story. This type of task can also be used by a single individual whose task it is to put the story together on his or her own. There are variations on this theme depending on one's research question. If, for example, one wanted to investigate unplanned speech, one could give participants the picture sequence and have them tell the story immediately. Alternatively, one could give participants time to think about the story and gather their thoughts. This would be appropriate if one were concerned with elements of planned speech.

3.7.3. Discourse Completion Test (DCT)

Perhaps the most common method of doing pragmatics-based research has been through the use of a DCT. This is particularly useful if one wants to investigate speech acts such as apologies, invitations, refusals, and so forth. One can manipulate relatively easily such factors as age differences or status differences between interlocutors. DCTs are implemented most frequently in writing, with the participants being given a description of a situation in which the speech act occurs. After the description, there is usually blank space where the response is required. The following example (Beebe & Takahashi,

1989, p. 109) illustrates a situation in which a status difference may be a factor when trying to provide embarrassing information to someone:

You are a corporate executive talking to your assistant. Your assistant, who will be greeting some important guests arriving soon, has some spinach in his/her teeth.

There are other instances when one needs to force a response. One way to do this is not only to provide space for the response, but to sandwich that space between the stimulus and the response to the response. For example, Beebe, Takahashi, and Uliss-Weltz (1990, p. 69) supplied the following discourse to elicit refusals:

Worker: As you know, I've been here just a little over a year now, and I know you've been pleased with my work. I really enjoy working here, but to be quite honest, I really need an increase in pay.

Worker: Then I guess I'll have to look for another job.

One can also ask for judgments of appropriateness following a description of scene-setting information, as in the following example:

Yesterday everything went badly. You were flying from Dayton, Ohio, to New York for a job interview. You were pleased because you were one of the final candidates. On your way to the airport, there was a water main break and the highway was flooded, which caused a closure of the highway. You had to take back roads to the airport (an area of town with which you were not familiar), but arrived too late for your flight. You were going to call the personnel manager to tell her of your predicament, but you couldn't find a phone. Just then you realized that there was another plane to New York that would still get you there in time. You boarded the plane, but because of storms in the New York area, your plane circled and circled the airport. When you landed, you were late for your appointment. The office was closed and you had to wait until this morning to talk to the personnel manager.

What will you say when you speak with her?

1. I would like to take this opportunity to apologize for missing the scheduled meeting. I'm sure I'll never let you down again. ___yes ___no
2. I would like you to give me another chance. ___yes ___no
3. I'm sorry that I didn't call earlier, but I was tired and so I slept late. ___yes ___no
4. I really, really want to work in your company. I want to make good use of my studies. ___yes ___no
5. I sincerely apologize for not making the interview. Because of the storms, my plane circled the airport for over an hour and I couldn't call you. We didn't land until after 5:00. I would appreciate it if I could reschedule my interview. ___yes ___no

Judgments can be dichotomous, as in the previous example, or they could be scalar. Alternatively, the situation could be presented with instructions to the nonnative speaker to state what her or she would say to the personnel manager. A word of caution is in order. The responses represent what a learner believes he or she would say in a particular context. This may or may not correspond to what would actually be said. Thus, results such as these need to be interpreted cautiously or at least verified against real situations whenever possible.

3.7.4. Role Play

In general, there are two types of role plays: open and closed. Closed role plays are similar to discourse completion tasks but in an oral mode. Participants are presented with a situation and are asked to give a one-turn oral response. Open role plays, on the other hand, involve interaction played out by two or more individuals in response to a particular situation. The limits that are given in closed role plays are not present to any significant degree in open role plays. Closed role plays suffer from the possibility of not being a reflection of naturally occurring data. Open role plays reflect natural data more exactly, although one must recognize that they are still collected in a nonnatural environment and thus are subject to the same difficulties as are closed role plays (see Gass & Houck, 1999, for an example of a study using an open role play).

3.7.5. Video Playback for Interpretation

In pragmatics research we might be interested in how people react to pragmatic infelicities. For example, how might the native speaker profes-

sor have reacted in the following situation (from Goldschmidt, 1996, p. 255), and why?

NNS: I have a favor to ask you.

NS: Sure, what can I do for you?

NNS: You need to write a recommendation for me.

In this particular case, a researcher might have asked the professor how he or she interpreted this somewhat abrupt request for a letter of recommendation. It might also be interesting to investigate this issue further by varying the context (a nonnative speaker professor, a native speaker professor, etc.). One way to accomplish this would be to stage scenarios according to variables of interest, videotape them, and prepare specific questions for observers. Bardovi-Harlig and Dörnyei (1998) attempted to determine reactions to pragmatic and grammatical errors by videotaping staged clips of nonnative speakers making them. Listeners (ESL and EFL learners and teachers) were given the following questionnaire and asked to rate each episode (p. 244):

Stimulus from video: I'm really sorry but I was in such a rush this morning and I didn't brought it today.

Was the last part appropriate / correct? Yes No

If there was a problem, how bad do you think it was?

Not bad at all :___:___:___:___:___ Very bad

3.8. QUESTIONNAIRES AND SURVEYS

Brown (2001) defined questionnaires (a subset of survey research) as "any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting them among existing answers" (p. 6). The survey, typically in the form of a questionnaire, is one of the most common methods of collecting data on attitudes and opinions from a large group of participants; as such, it has been used to investigate a wide variety of questions in second language research. Questionnaires allow researchers to gather information that learners are able to report about themselves, such as their beliefs and moti-

variations about learning or their reactions to learning and classroom instruction and activities—information that is typically not available from production data alone.

Specialized types of questionnaires have also been developed to address specific research areas or questions. For example, as noted previously, discourse completion questionnaires have been used to investigate interlanguage pragmatics.

In addition to different varieties of questionnaires, two types of questionnaire items may be identified: closed and open ended. A closed-item question is one for which the researcher determines the possible answers, whereas an open-ended question allows respondents to answer in any manner they see fit. Closed-item questions typically involve a greater uniformity of measurement and therefore greater reliability. They also lead to answers that can be easily quantified and analyzed. Open-ended items, on the other hand, allow respondents to express their own thoughts and ideas in their own manner, and thus may result in more unexpected and insightful data. An example of a closed-item question is, “How many hours a week did you study to pass this test? Circle one: 3, 4, 5, or 6 or more.” An example of a more open-ended question is, “Describe ways that you found to be successful in learning a second language?”

The type of questions asked on a questionnaire naturally depends on the research questions being addressed in the study. For example, in relatively unstructured research, it may be more appropriate to ask open-ended questions and allow participant responses to guide hypothesis formation. Once hypotheses are formulated, researchers can ask closed-item questions to focus in on important concepts. Of course, questionnaires need not be solely closed or open ended, but can blend different question types depending on the purpose of the research and on what has previously been learned about the research phenomenon. For a more in-depth discussion of these considerations, as well as a practical guide for the use of questionnaires in second language research, see Dörnyei’s (2003) text, which provided a helpful list of published questionnaires illustrating the sheer range of research that has been carried out using this approach, as noted in the following list:

Dörnyei’s (2003, pp. 144-149) text on questionnaires includes second language research on the following topics:

-
- | | | |
|------------------------------|--|------------------------------|
| • Language attitudes | • Biographic background | • Classroom observation |
| • Computer familiarity | • feedback | • Group cohesiveness |
| • Immigrant settlement | • Language anxiety | • Language contact |
| • L2 course evaluation | • L2 learner beliefs | • L2 learning motivation |
| • L2 learning strategies | • L2 learning styles | • Linguistic self-confidence |
| • Needs analysis | • Self-evaluation | • Teacher anxiety |
| • Teacher beliefs | • Teacher evaluation | • Teacher motivation |
| • Teacher self-evaluation | • Preferences for instructional activities | |
| • Willingness to communicate | | |
-

Questionnaires can also be customized. An example of a highly customized elicitation procedure of this nature is a grid-based scheme. A researcher creates a grid following analysis of a completed questionnaire, and/or carrying out an in-depth interview with the participant. The grid is designed to both reflect the participants' input and uncover further information, including their perceptions about the patterns and relationships in the data collected to date. For example, drawing on work based on teachers of mathematics and sciences, Breen, Hird, Milton, Oliver, and Thwaite (2001) created grids to uncover information about teachers' principles and classroom practices. An example of one of their grids appears in Fig. 3.6. As Breen et al. noted:

[P]rior to the second interview, the researcher drew up a grid for each individual teacher, transcribing the teacher's descriptions of practices and their reasons for them from the cards. The teacher's practices were listed on the vertical axis and their reasons listed on the horizontal axis ... at this second interview the researcher worked with the teacher on the grid eliciting information as to whether the teacher saw a relationship between each action in turn, and all the reasons on the vertical axis. (pp. 478–479)

One of the primary advantages of using questionnaires is that, in addition to being more economical and practical than individual interviews, questionnaires can in many cases elicit longitudinal information from learners in a short period of time. Questionnaires can also elicit comparable information from a number of respondents. In addition, questionnaires can be administered in many forms, including via e-mail, by phone, through mail-in forms, as well as

Reasons for actions					
	Quieter students should have a chance to speak/use the language	Students should do things they like doing if they think it's useful for learning	Student's contributions are important for confidence and motivation	Need to simplify or break down the task to give a sense of progress	Build on what students already know
Actions					
Begins the lesson by revising work from previous lesson	2	2	2	1	1
Accepts and encourages students' spontaneous suggestions	1	1	1	3	1
Encourages students to write down new items of language	3	3	3	1	3
Gets students to highlight words on the handout	3	3	3	1	3
Expects students to speak in English in the pair work	1	3	3	1	1
Gets some students to sound out individual words that are new	1	1	3	1	1

FIG. 3.6. Sample grid. Source: Breen, M. P., Hird, B., Milton, M., Oliver, R., & Thwaite, A. (2001). Making sense of language teaching: Teachers' principles and classroom practices. *Applied Linguistics*, 22(4), 479. Copyright © 2001 by Oxford University Press. Reprinted with the permission of Oxford University Press.

in person, allowing the researcher a greater degree of flexibility in the data gathering process. Depending on how they are structured, questionnaires can provide both qualitative insights and quantifiable data, and thus are flexible enough to be used in a range of research.

There are potential problems related to the analysis of questionnaire data. One concern is that responses may be inaccurate or incomplete because of the difficulty involved in describing learner-internal phenomena such as perceptions and attitudes, for example. This may be the case if the questionnaire is completed in the L2, in which lower proficiency in the L2 may constrain the answers. Both learners and native speakers might be able to provide salient details, but they may not be able to paint a complete picture of the research phenomenon. This being so, questionnaires usually do not provide a complete picture of the complexities of individual contexts. This is especially important to remember when using open-ended written questionnaires, because participants may be uncomfortable expressing themselves in writing and may choose to provide abbreviated, rather than elaborative, responses. Hence, whenever possible, questionnaires should be administered in the learners' native language, learners should be given ample time to specify their answers, and learners with limited literacy should be given the option of providing oral answers to the questionnaire (which can be recorded).

Another concern is that even though it is often assumed that researchers can control or eliminate bias by using questionnaires, it is also possible, as with any type of elicitation device, that the data elicited will be an artifact of the device. Thus, for example, if a study utilizes a discourse completion questionnaire, the researcher should take particular caution when interpreting the results, because the situations depicted are usually hypothetical. In this type of questionnaire, learners are only indicating how they think they would respond; this may or may not correspond to how they would actually respond in real life.

To maximize the effectiveness of the questionnaire, researchers should try to achieve the following:

- Simple, uncluttered formats.
- Unambiguous, answerable questions.
- Review by several researchers.
- Piloting among a representative sample of the research population.

This should be done before undertaking the main bulk of data collection to ensure that the format is user-friendly and the questions are clear.

3.9. EXISTING DATABASES

In addition to the many elicitation techniques we have discussed in this chapter, there are existing databases as well, consisting of data that have already been collected, transcribed, and often analyzed. If the research questions allow it, using an existing database can save considerable amounts of time and effort. The main database for language acquisition research is CHILDES (MacWhinney, 2000), which focuses on spoken language. Other databases include corpora dealing with various aspects of performance (e.g., writing), whereas still others are available only in languages other than English and serve very specific purposes. We are unable to deal with these here, but encourage those who want to do second language research in a specific language to learn about them. Web-based searches are a good place to start, using particular parameters of relevance to your research. We now turn to a short description of CHILDES.

3.9.1. CHILDES

The CHILDES database was designed to facilitate language acquisition research. It allows researchers to study conversational interactions among child and adult first and second language learners and includes a variety of languages and situations/ contexts of acquisition, including bilingual and disordered acquisition, as well as cross-linguistic samples of narratives. It consists of three main components: CHILDES, a database of transcribed data; CHAT, guidelines for transcription and methods of linguistic coding in a format in which users can create “dependent tiers” (in addition to a main tier containing speakers’ utterances) to record their notes on context, semantics, morphology, and syntax; and CLAN, software programs for analyzing the transcripts through, for example, searches and frequency counts. It is also possible in CHILDES to link transcripts to digital audio and video recordings.

3.9.2. Other Corpora¹⁴

Granger (2002) provided a review of the ways in which corpora can be used in second language research, along with numerous references to existing corpora. In general, according to Granger, “[C]omputer learner corpora

¹⁴The following website (<http://calper.la.psu.edu/corpus.php>) provides resources on learner corpora from a wide variety of languages.

are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance” (p. 7).

An important consideration in using corpora is to understand where the data come from and how the data are organized. For example, if one is using a corpus based on written essays, one must have access to basic information such as the prompt that was used to elicit the essay. This is necessary if one plans to make any sort of comparison across languages or across times.

Corpora can be organized in different ways. For example, some might be used to study idioms or collocations, and the database might be tagged for that purpose. Others may be tagged for errors and/or parts of speech. This could be useful if one were looking at German word order acquisition, for example, and wanted to know how many instances could be found of verbs in second position in sentences whose first word is not a noun or pronoun.

As noted previously, corpora can be useful, but a clear understanding of what is and what is not included is essential to an understanding of how they can be used appropriately.

3.10. CONCLUSION

This chapter has provided some preliminary information on some of the more common methodological tools used for eliciting data in second language research. This is by no means an exhaustive list, but was intended to acquaint the reader with some of the issues surrounding each instrument. As we have repeatedly noted, with materials and methods one must pilot the instruments to ensure that they elicit what one intends and to ensure their appropriateness for the study. In the next chapter, we focus on the actual design of a study. In particular, we look at different types of research variables, together with the important concepts of validity and reliability.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Take a research question that you came up with from chapter 1 (or another research question in which you are interested). What elicitation tool(s) would you use to investigate it, and why? Are there any alternatives? If so, which would you choose, and why?
2. Conduct a library search. Find three articles that investigate *similar* topics. Do they use the same elicitation tool? If so, why do you think

- this is the case (e.g., it is the only possibility)? Do you think that there could have been an alternative? If so, describe the alternative and how it might be better or worse than the one originally used.
3. Conduct a library search. Find three articles that investigate *different* topics. Do they use different elicitation tools? Could they have used other elicitation tools? If not, why not? If so, come up with some other means of eliciting the type of data that they need in order to answer the research questions that they have set out.
 4. Find two recent articles in a second language journal that could have dealt with unanswered questions through a recall procedure. Describe how you would have conducted a recall portion for the study that you have selected.
 5. You want to determine the extent to which learners understand aspects of complex syntax, focusing in particular on to what the noun phrases are referring. You have decided to use sentences like the following and have elected to use elicited imitation. What factors do you need to take into account? Come up with six test sentences and describe what you will do. Besides elicited imitation, how could you elicit information regarding the appropriate meaning of the pronoun?
 - a. When he entered the office, the janitor questioned the man.
 - b. As she walked to the blue door, Anne wondered about Joan's father.
 6. You want to elicit speech samples containing:
 - a. Subjunctive (*I request that everyone be here by 5.*).
 - b. Embedded questions (*The teacher asked why she was late.*).How might you go about doing this?
 7. Assume that you want to investigate how native speakers react to requests by a second language speaker. Further assume that you believe that it is not so much the words people use that affect different native speaker reactions, but instead the stereotypes that native speakers have formed about particular groups of nonnative speakers (e.g., French speakers). How would you go about investigating this?
 8. You want to study agreement in English (e.g., subject-verb agreement). How would you go about collecting data?
 9. Describe the major benefits of conducting a pilot test on materials.

Research Variables, Validity, and Reliability

This chapter focuses on the concepts necessary for understanding how to design a study in second language research. We begin with an outline of variables and scales, and follow with descriptions of specific types of validity and reliability. We also discuss sampling, representativeness and generalizability, and the collection of biographical data.

4.1. INTRODUCTION

In chapter 1, we introduced the concepts of research questions and research hypotheses. Research questions can take a range of forms. One example of a specific and answerable research question might be, “What is the effect of form-focused instruction on the acquisition of English relative clauses by French- and Japanese-speaking learners of English?” Because of differences between Japanese and English and similarities between French and English, we might hypothesize as follows: “French-speaking learners of English will perform better following form-focused instruction than will Japanese-speaking learners of English.” Assuming that the research question is clearly phrased, answerable, and motivated by the literature, we can move on to the research hypotheses.

4.2. HYPOTHESES

A hypothesis is a type of prediction found in many experimental studies; it is a statement about what we expect to happen in a study. In research reports there are generally two types of hypotheses: research hypotheses and null hypotheses. The null hypothesis (often written as H_0) is a neutral statement used as a basis for testing. The null hypothesis states that there is no re-

relationship between items under investigation. The statistical task is to reject the null hypothesis and to show that there is a relationship between X and Y. Given our hypothesis above that French-speaking learners of English would perform better following form-focused instruction than would Japanese-speaking learners of English, the null hypothesis would be:

There will be no difference between the performance of the French group and the Japanese group on a posttest.

We could then statistically test the differences in performance between these groups on a posttest following instruction to determine if any differences found were due to chance or due to treatment. We return to hypotheses and statistics in chapter 9.

When, based on previous research reports in the literature, we expect a particular outcome, we can form research hypotheses. There are two ways that we can do this. The first is to predict that there will be a difference between two groups, although we do not have sufficient information to predict the direction of the difference. For example, we might have a research hypothesis that states simply that the two groups will be different, such as:

There will be a difference between the performance of the French-speaking group and the Japanese-speaking group on a posttest.

This is known as a nondirectional or two-way hypothesis.

On the other hand, we may have enough information to predict a difference in one direction or another. This is called a directional or one-way hypothesis. To continue our example, we might believe (based on the closer linguistic relationship between English and French than between English and Japanese) that the French-speaking group will perform better than the Japanese-speaking group. We would then formulate our hypothesis as follows:

The French-speaking group will perform better on a posttest than the Japanese-speaking group.

4.3. VARIABLE TYPES

In order to carry out any sort of measurement, we need to think about variables; that is, characteristics that vary from person to person, text to text, or object to object. Simply put, variables are features or qualities that change.

TABLE 4.1
Variable Types

<i>Research Question</i>	<i>Independent Variable</i>	<i>Dependent Variable</i>
Does feedback type affect subsequent performance?	Feedback type	Performance measure
Can elements of child-directed speech aid in learning morphology?	Child-directed speech	Measure of morphological acquisition
Does length of residence affect identification of word-final consonants?	Length of residence	Measure of success in identifying word-final consonants
Is there a relationship between learners' noticing of recasts and L2 development?	Noticing of recasts	L2 development measure

For example, we might want to think about the effects of a particular pedagogical treatment on different groups of people (e.g., Spanish speakers learning English versus Japanese speakers learning English). Native language background, then, is a variable. What we are ultimately doing in experimental research is exploring whether there are relationships between variables and a constant:

Example:

We want to examine the effects of different types of instruction on a group of foreign language learners. We take as our object of investigation students enrolled in first semester foreign language classes of Spanish. We use two equivalent classes of first semester Spanish (selecting classes is an issue that we discuss in chap. 5). We have teachers provide one group with explicit grammar instruction. Another group receives no grammar instruction, but receives a significant amount of input on the specific linguistic structure in question.

The variable under investigation: Type of instruction.

What is being held constant: Class level (first semester Spanish, native language background of participants).

4.3.1. Independent and Dependent Variables

There are two main variable types: independent and dependent. The independent variable is the one that we believe may “cause” the results; the dependent variable is the one we measure to see the effects the independent variable has on it. Let us consider the examples in Table 4.1.

In each of the examples in Table 4.1, the independent variable is manipulated to determine its effect on the dependent variable. To elaborate on one of these, let us consider the third example: Does length of residence affect identification of word-final consonants? Let us assume that we have a well-motivated reason to believe that learners are able to recognize word-final consonants based on the amount of input to which they have been exposed. Assuming that one can operationalize (see sec. 4.4.) amount of input (possibly as length of residence in a foreign country or amount of classroom exposure), we would then divide learners into groups depending on their exposure to the target language and see if there is a difference in the degree of correct identification of word-final consonants between or among the groups. The dependent variable would be expressed in terms of the number or percentage of word-final consonants correctly identified, and we would determine whether learners with longer or greater exposure (independent variable) had higher scores (dependent variable).

It is clear that the variables in Table 4.1 also differ in another way—some can be directly manipulated by the researcher (e.g., feedback types), whereas some already exist (e.g., amount of input). With those that exist, the researcher needs to find the right way of selecting the appropriate forum for investigating the effects. With those that can be manipulated, the task of the researcher is to determine how to manipulate the variable appropriately. For example, in the case of feedback types, one could select three different teachers who naturally employ different feedback types and use their classrooms for investigation. Alternatively, one could train teachers to use different feedback types.

4.3.2. Moderator Variables

Moderator variables are characteristics of individuals or of treatment variables that may result in an interaction between an independent variable and other variables. Let us assume again a study on the effect of length of residence on the recognition of word-final consonants. Let us further assume that we have a theoretical rationale for believing that

length of residence might differentially affect recognition depending on gender. Gender might then be considered to be a moderator variable. In other words, a moderator variable is a type of independent variable that may not be the main focus of the study, but may modify the relationship between the independent variable and the dependent variable. Of course, moderator variables can “sneak” into a study without the researcher realizing that they may be important. One could imagine that in the hypothetical study mentioned earlier, it might not occur to the researcher that gender would be a factor. We have to be cognizant, therefore, of the fact that there may be variables that interfere with the actual results we are seeking. These are known as intervening variables.

4.3.3. Intervening Variables

Intervening variables are similar to moderator variables, but they are not included in an original study either because the researcher has not considered the possibility of their effect or because they cannot be identified in a precise way. For instance, consider a study that measures the effect of pedagogical treatment (independent variable) on learners’ overall language proficiency (dependent variable, as measured by TOEFL scores). A variable that cannot be measured or understood easily might be the individuals’ test-taking abilities. In other words, the results may be due to test-taking abilities rather than to the treatment. Because this variable was not controlled for, it is an intervening variable that could complicate the interpretation of the results.

4.3.4. Control Variables

When conducting research, one ideally wants to study simply the effects of the independent variable on a dependent variable. For example, consider the impact of feedback type on a performance measure. Variables that might interfere with the findings include the possibility that learners with different levels of proficiency respond differently to different types of feedback. Another possibility is that different students, depending on their prior language learning experiences, respond differently to different types of feedback. Whenever possible, researchers need to identify these possible factors and control for them in some way, although it should be recognized that identifying and controlling for all variables in L2 research may be difficult.

One way to determine if gender or possibly native language background (as a way of operationalizing language learning experiences) might have an

effect is to balance these variables by having an equal number of men versus women or an equal number of Korean versus Japanese versus Spanish speakers (or whatever languages one is dealing with). These then become moderator variables (see earlier discussion). Another way to control for possibly interfering, or confounding, variables is to eliminate the variable completely (i.e., to keep it constant). In our hypothetical example, our study might include only men or only women, or only Korean or Japanese or Spanish speakers. Gender and native language then become control variables. This latter solution, of course, limits the degree of generalizability of one's study (see sec. 4.6.7 on external validity).

4.4. OPERATIONALIZATION

In many instances in second language research it is difficult to measure variables directly so researchers provide working definitions of variables, known as operationalizations. An operational definition allows researchers to operate, or work, with the variables. Operationalizations allow measurement. To return to the earlier example, we said that we need to operationalize "amount of input," because this term, as stated, is vague. Although it might be difficult to come up with a uniform concept of amount of input, it is possible to think of examples in which groups vary along some parameter that seems close to the amount of input. For example, classroom learners could be classified based on how many years of exposure they have had to the target language, for example 1 versus 2 versus 3 years. Hence, "amount of input" could be operationalized as years of exposure in this case. In a more natural setting, the operationalization of "amount of input" could be the number of years spent in the target language environment. Once a variable has been operationalized in a manner such as this, it is possible to use it in measurements.

4.5. MEASURING VARIABLES: SCALES OF MEASUREMENT

We have discussed variable types, but there is another way that we can think of differences in variables: What scales are going to be used to describe and analyze the data from the different variables? In this section, we provide a brief introduction to different scales. Chapter 8 on data coding deals with the topic in greater detail.

The three most commonly used scales are nominal, ordinal, and interval. Ratio scales, a type of interval scale, are not included here because they

are not used as frequently in the type of research that is carried out in second language studies.¹ Nominal scales are used for attributes or categories and allow researchers to categorize variables into two or more groups. With nominal scales, different categories can be assigned numerical values. For example, in a study of gender, (1) may be assigned to male and (2) to female. The numbers indicate only category membership; there is no indication of order or magnitude of differences. Consequently, in a nominal scale the concept of average does not apply.

An ordinal scale is one in which ordering is implied. For example, student test scores are often ordered from best to worst or worst to best, with the result that there is a 1st-ranked student, a 2nd-ranked student, a 10th-ranked student, and so forth. Although the scores are ordered, there is no implication of an equal distance between each rank order. Thus, the difference between Students 1 and 2 may not be the same as the difference between Students 2 and 3. It is also often the case that researchers need to give holistic judgments to student work. This might be the case, for example, with second language writing scores. If we gave writing scores on a scale from 1 to 100, we might not be able to say that someone who received an 80 is twice as good a writer as someone who received a 40 without having precise information about what 40 and 80 meant on the scale. An ordinal scale might be useful in ordering students for placement into a writing program, but we cannot make judgments about exactly how much better one student is than another.

An interval scale represents the order of a variable's values, but unlike an ordinal scale it also reflects the interval or distance between points in the ranking. If a test represents an interval scale, then one can assume that the distance between a score of 70 and 80 is the same as the distance between 80 and 90. Thus, we could say, for example, that someone who received a score of 10 on a vocabulary test knew twice as many of the words that were tested as did someone who received a 5. As this example shows, an interval scale implies measurable units, such as number of correct answers, years of residence in the target language country, or age (as opposed to being a good writer).

4.6. VALIDITY

After spending a great deal of time and effort designing a study, we want to make sure that the results of our study are valid. That is, we want them to reflect what we believe they reflect and that they are meaningful in the

¹Ratio scales have a true zero point where zero represents the absence of the category.

sense that they have significance not only to the population that was tested, but, at least for most experimental research, to a broader, relevant population. There are many types of validity, including content, face, construct, criterion-related, and predictive validity. We deal with each of these in turn before turning to internal and external validity, which are the most common areas of concern.

4.6.1. Content Validity

Content validity refers to the representativeness of our measurement regarding the phenomenon about which we want information. If we are interested in the acquisition of relative clauses in general and plan to present learners with an acceptability judgment task, we need to make sure that all relative clause types are included. For example, if our test consists only of sentences such as “The boy who is running is my friend,” we do not have content validity because we have not included other relative clause types such as “The dog that the boy loves is beautiful.” In the first sentence the relative pronoun *who* is the subject of its clause, whereas in the second sentence the relative pronoun *that* is the object. Thus, our testing instrument is not sensitive to the full range of relative clause types, and we can say that it lacks content validity.

4.6.2. Face Validity

Face validity is closely related to the notion of content validity and refers to the familiarity of our instrument and how easy it is to convince others that there is content validity to it. If, for example, learners are presented with reasoning tasks to carry out in an experiment and are already familiar with these sorts of tasks because they have carried them out in their classrooms, we can say that the task has face validity for the learners. Face validity thus hinges on the participants’ perceptions of the research treatments and tests. If the participants do not perceive a connection between the research activities and other educational or second language activities, they may be less likely to take the experiment seriously.

4.6.3. Construct Validity

This is perhaps the most complex of the validity types discussed so far. Construct validity is an essential topic in second language acquisition research precisely because many of the variables investigated are not easily or di-

rectly defined. In second language research, variables such as language proficiency, aptitude, exposure to input, and linguistic representations are of interest. However, these constructs are not directly measurable in the way that height, weight, or age are. In research, construct validity refers to the degree to which the research adequately captures the construct of interest. Construct validity can be enhanced when multiple estimates of a construct are used. For example, in the hypothetical study discussed earlier that was seeking to link exposure to input with accuracy in identifying final consonants, the construct validity of the measurement of “amount of input” might be enhanced if multiple factors—such as length of residence, language instruction, and the language used in the participants’ formal education—were considered together.

4.6.4. Criterion-Related Validity

Criterion-related validity refers to the extent to which tests used in a research study are comparable to other well-established tests of the construct in question. For example, many language programs attempt to measure global proficiency either for placement into their own program or to determine the extent to which a student might meet a particular language requirement. For the sake of convenience, these programs often develop their own internal tests, but there may be little external evidence that these tests are measuring what the programs assume they are measuring. One could measure the performance of a group of students on the local test and a well-established test (e.g., TOEFL in the case of English, or in the case of other languages, another recognized standard test). Should there be a good correlation (see chap. 9 for a discussion of correlations in statistics), one can then say that the local test has been demonstrated to have criterion-related validity.

4.6.5. Predictive Validity

Predictive validity deals with the use that one might eventually want to make of a particular measure. Does it predict performance on some other measure? Considering the earlier example of a local language test, if the test predicts performance on some other dimension (class grades), the test can be said to have predictive validity.

We now turn to the two main types of validity that are important in conducting research: internal validity and external validity.

4.6.6. Internal Validity

Internal validity refers to the extent to which the results of a study are a function of the factor that the researcher intends. In other words, to what extent are the differences that have been found for the dependent variable directly related to the independent variable? A researcher must control for (i.e., rule out) all other possible factors that could potentially account for the results. For example, if we wanted to observe reaction times to a set of grammatical and ungrammatical sentences, we might devise a computer program that presents sentences on a computer screen one at a time, with learners responding to the acceptability/unacceptability of each sentence by pressing a button on the computer. To make the task easier for the participants in the study, we could tape the letter *A* for “acceptable” over the letter *t* on the keyboard and tape the letter *U* for “unacceptable” over the *y* key on the keyboard. After we have completed the study, someone might ask us if we checked for handedness of the participants. In other words, could it be the case that for those who are left handed, the *A* key (“acceptable”) might be faster not because it is faster to respond to acceptable as opposed to unacceptable sentences (part of our hypothesis), but because left hands on left-handed people react faster. Our results would then have been compromised. We would have to conclude that there was little internal validity.

It is important to think through a design carefully to eliminate or at least minimize threats to internal validity. There are many ways that internal validity can be compromised, some of the most common and important of which include participant characteristics, participant mortality (dropout rate), participant inattention and attitude, participant maturation, data collection (location and collector), and instrumentation and test effects.

4.6.6.1. Participant Characteristics

The example provided in the previous section concerning handedness is a participant characteristic. Clearly, not all elicitation techniques will require controlling for handedness. In other words, there may be elements of the research questions and/or elicitation technique that require a careful selection of one characteristic or another. Let us consider some relevant participant characteristics for second language research: language background, language learning experience, and proficiency level.

Language Background. In many studies, researchers want to compare one group of students with another group based on different treat-

ments. For example, let us assume that a study on the role of attention in second language learning compared groups of students in a foreign language class who were exposed to a language structure with and without devices to ensure that they paid attention to that structure. It would be important that each group of students be relatively homogeneous. Were they not homogeneous, one could not be sure about the source of the results. For instance, let's further assume that one group of students had a large number of participants who were familiar with a language closely related to the target language (either through exposure at home or in the classroom). We then could not distinguish between the effects of the treatment and the effects of the background knowledge of the participants.

Language Learning Experience. Participants come to a language learning situation with a wide range of past experiences. In some instances, these experiences may have importance for research. For example, many students in an ESL setting have had prior English instruction in their home countries, and this prior instruction may differ from one country to another. If we wanted to conduct a study in which we compared implicit versus explicit methods of instruction, we might find that a group that received explicit instruction outperformed a group that received implicit instruction. If the two groups also differed in terms of prior learning experiences, we would be left with two variables: learning experience and instruction type. We would not be able to distinguish between them. Are our results possibly due to the fact that explicit instruction yielded a better outcome because one group was more familiar with and thus more affected by that type of instruction? Or did one instruction type yield better results due to that type of instruction? It is the latter that we want our study to measure.

Proficiency Level. This is one of the most difficult areas to control for when conducting second language research. In foreign language environments, the issue is perhaps simpler than in second language environments, because in the former but not the latter there is limited exposure outside the classroom although here, too, there can be problems. In the area of foreign language research, there are some global proficiency measures such as the Oral Proficiency Interview (OPI) so that learners can be matched for proficiency. Another common measure is to use placement in class level (first year versus second year versus third year, etc.). In a foreign language environment, this is relatively "safe" because exposure is more or

less limited to what occurs in the classroom.² However, with second language learners, backgrounds and outside experiences are varied and there is typically unevenness in skill levels. For example, some potential participants in the same class level may have excellent oral skills but weak written skills, and vice versa. It is therefore important to consider how this may bear on the specific research questions of the study.

We have discussed some of the ways in which participant characteristics differ. It is also important to ensure that participants are matched on the feature that is being examined. For example, if one is conducting a study that investigates the perception and production of phonological categories, it may not be sufficient to assume that advanced students are better than intermediate students because the intermediate students may have spent more time in the country where the language is spoken than the advanced students and, consequently, their perception and production of target language sounds may be more advanced even if their command of other aspects of the language is not. One must also be wary of using global proficiency tests when the testing instrument relies on one skill or another. For example, a global language test that provides information on grammar and vocabulary may obscure differences in participants' listening abilities. If listening is a major part of gathering data (e.g., as in elicited imitation tasks, discussed in chap. 3), an additional measure of listening ability may be needed to make sure that difficulty with the instrument is not an issue causing problems with internal validity.

4.6.6.2. Participant Mortality

Some studies that are conducted in second language research are longitudinal in nature. That is, they seek to measure language development by sampling over time. As such, researchers may typically carry out immediate posttests and also one or more delayed posttests to determine the shorter- and longer-term effects of a treatment. In order to appropriately address research questions, it is best to ensure that all participants are present for all sessions. However, in many classroom research settings, it is inevitable that not all participants will be present at all times. A researcher must determine how to deal with this situation, and there a number of factors

²This is, of course, an oversimplification, because classroom learners will vary greatly in terms of the amount of time they spend out of class reading the foreign language or in the language laboratory.

that one might want to consider. For example, if a researcher has 50 participants and one of them has to be eliminated, the loss is probably not significant. If, on the other hand, participant numbers are balanced across groups, the loss of a participant in one group may necessitate the elimination of a matched participant in another group. Some possible scenarios follow.

Scenario 1: Participant Missing From One Treatment Session

<i>Purpose of study:</i>	Measuring the effect of quantities of input across groups.
<i>Number of participants:</i>	25 per group.
<i>Method:</i>	Differing amounts of input per lesson; five lessons over a 2-week period.
<i>Posttests:</i>	One posttest.
<i>Situation:</i>	One student in one group misses one class period.
<i>Issue:</i>	Should the posttest data for that student be included in the final data pool?
<i>Response:</i>	It might depend on how the groups vary in terms of input. Given that this study is measuring quantities of input and that one group may vary from others by only small differences in the amount of input, the inclusion of someone who missed one class session might make him or her more like someone in another group. Thus, data from this learner should probably be eliminated.

Scenario 2: Participant Missing From One Posttest

<i>Purpose of study:</i>	Determining the long-term effects of attention.
<i>Number of participants:</i>	Two groups of 10 each.
<i>Method:</i>	Computer-based input varying attention conditions.
<i>Posttests:</i>	Five posttests given at 1-month intervals.
<i>Situation:</i>	One student in one group misses one posttest.
<i>Issue:</i>	Should the data for that student be included in the final data pool?
<i>Response:</i>	Given that there are five post-tests, one could make a decision in advance that a student must

participate in at least the first and the last of the posttests and two of the remaining three. This would allow some flexibility in keeping as many participants in the data pool as possible while still providing the researcher with information from four data points following the treatment.

Scenario 3: Participant Missing From Part of Posttest

<i>Purpose of study:</i>	Determining the long-term effects of attention on syntax versus vocabulary.
<i>Number of participants:</i>	Two groups of 10 each.
<i>Method:</i>	Computer-based input varying attention conditions.
<i>Posttests:</i>	One posttest for syntax and one for vocabulary.
<i>Situation:</i>	One student in one group misses one posttest (either syntax or vocabulary).
<i>Issue:</i>	Should the data for that student be included in the final data pool?
<i>Response:</i>	Given that there are two separate posttests and assuming that data are being aggregated rather than each student's performance on syntax being compared to his or her performance on vocabulary, one could maintain the data in the pool. This would mean that data for the statistical tests would include 10 syntax scores versus 9 vocabulary scores (or vice versa). If, on the other hand, one wanted to do a comparison of each person's data on the two tasks, then the individual who missed one would not be able to be kept in the final data pool.

We have presented three sample scenarios showing what some of the considerations might be in determining what to do about participant mortality. Each situation will, undoubtedly, be different and will require a different (and justifiable) solution. The point to remember is the importance of carefully thinking through the various possibilities given the design of the experiment or longitudinal sessions, and making a principled decision as to how to solve the problem in the event of participant absences. These decisions should not be made ad hoc; when possible, they

should be made in advance of data collection. They should also be fully reported in the research report.

4.6.6.3. Participant Inattention and Attitude

When we collect data from participants, we usually make the assumption that they are giving us their “best effort.” In other words, we rely on the notion that the language data we are collecting are uncontaminated by the experiment itself. This may not always be true. One factor that might affect participant behavior is what is known as the Hawthorne effect, which refers to the positive impact that may occur simply because participants know that they are part of an experiment and are, therefore, “different” from others. Participants may also try to please the researcher by giving the answers or responses they think are expected. This is known as the halo effect. Hawthorne and halo effects are also discussed in chapters 6 and 7 in relation to experimental designs and qualitative research.

Participating in a study also has potential negative effects. For example, researchers might want to consider factors such as fatigue and boredom when asking participants to perform tasks. This was mentioned in chapter 3 in the discussion of the number of sentences to use in an acceptability judgment test. Whatever method is being used to gather data, one needs to think of the exhaustion and boredom factor. How much time can one reasonably ask a participant to perform without losing confidence in the results, especially if it is a repetitive and demanding task such as judging sentences? There is no magic answer; we must weigh the need to gather sufficient data against these factors. As discussed earlier, presenting tasks or items in different orders can serve to balance these effects.

A second factor is general inattentiveness, whether from the outset of the experiment or as a result of the experiment. In a study by Gass (1994) that involved giving participants the same task after a 1-week interval, the author noted that some participants provided diametrically opposed responses at the two time periods. In a stimulated recall after the second session, one of the participants stated that his results from the two sessions differed because his mind was wandering given that he had two academic tests that week. Clearly, one does not always know whether this is an issue, but one needs to be aware of this as a possible way of explaining what may appear to be aberrant or divergent results. In general, if time and resources permit, it is helpful to do a stimulated recall with participants (possibly using the test measure as a stimulus) or a postexperiment interview or exit

questionnaire to ascertain if there might be extra-experimental factors that impacted learner responses or behaviors. Gathering such data from even a subset of participants can help in interpreting results.

4.6.6.4. Participant Maturation

Maturation is most relevant in longitudinal studies and particularly in those involving children. For example, a study that spans a year or longer will inevitably include participants who change in one way or another in addition to changes in language development. Adults may not change dramatically in a 1-year period, but children certainly do. Moreover, people who were comparable at the outset of the study may change in different ways due to different experiences over time. Thus, one must find a way to balance regular maturational factors against the requirements of the study. When maturation is a consideration, a control group not subjected to the treatment or intervention is appropriate wherever possible. The inclusion of a control group provides one way to test whether any changes occurred because of the experimental treatment or because of maturation.

4.6.6.5. Data Collection: Location and Collector

Not all research studies will be affected by the location of data collection, but some might. Some obvious concerns relate to the physical environment; for example, the environment for two groups given the same test might influence the results if one group is in a noisy or uncomfortable setting and the other is not. A perhaps less obvious effect of setting might occur in a study in which a researcher is trying to gather information from immigrant parents (perhaps through an oral interview) about their attitudes concerning their desires for their children to learn the target language. Informal interviews in their home might yield results that differ from those obtained in a formal school setting, where teachers in proximity could influence what the parents think they should say.

Another factor in some types of research relates to the person doing the data collection. Given the scenario mentioned earlier concerning families being surveyed about their attitudes toward their children's learning of the target language, one could imagine different results depending on whether or not the interviewer is a member of the native culture or speaks the native language.

4.6.6.6. Instrumentation and Test Effects

The test instrument is quite clearly an important part of many research studies. In this section we discuss three factors that may affect internal validity: equivalence between pre- and posttests, giving the goal of the study away, and test instructions and questions.

Equivalence Between Pre- and Posttests. One serious design issue relates to the comparability of tests. A difficult pretest with an easier post-test will make it more likely for improvement to be apparent after a treatment. The opposite scenario will make it more likely for no improvement to be apparent following a treatment. There are a number of ways to address comparability of tests. For example, when testing grammatical improvement following a treatment, one can keep the grammatical structure the same and change the lexical items. Doing this, however, requires ensuring comparable vocabulary difficulty. For example, the sentence *The dog ate the chair* does not involve the same vocabulary difficulty level as *The deer consumed the rhododendron*. One way to address this issue might involve consulting a word frequency index (e.g., Brown Corpus, Academic English, Academic Word List; see Francis & Kucera, 1982; Thorndike & Lorge, 1944) that lists words of the same frequency—that is, words that appear approximately the same number of times in a corpus of the same size and type.

Another way to ensure comparability is to establish a fixed group of sentences for all tests. If a set of 30 sentences were established, Participant A could have a random set of 15 of those on the pretest and the remaining 15 on the posttest. Participant B could also have a random set of 15 on the pretest and the remaining is on the posttest, but the two participants would in all likelihood not have the same sets of 15. This is quite easy to do on a computer, but it could be done without a computer as well, counterbalancing the test by giving half of a group one set of sentences on the pretest and the other set on the posttest and giving the sets of sentences to the other half of the group in the reverse order. This technique may also eliminate the possible practice effects or participant inattentiveness that might arise if learners were tested on the same set of sentences twice.

Another example of the importance of test comparability can be seen in conducting second language writing studies. Researchers need to be mindful of the need to choose appropriate topics about which to write. A pretest that is based on a compare and contrast essay might be quite different in structure and vocabulary than a posttest essay based on a topic of persuasion. It would not be meaningful to compare the two essays.

Giving the Goal of the Study Away. One of the problems in doing second language research is that one sometimes does not want participants to know the precise nature of the language area or behavior that is being tested. We might want to conceal the precise nature of the study because we want responses that reflect natural behavior rather than what participants think they should say or do (see chap. 2 for a discussion of consent forms and how to strike a balance between not being deceptive and yet not revealing precisely what the study's focus is). This becomes particularly problematic when using a pretest because the pretest may in and of itself alert participants to the study's objective. One way of avoiding this problem is by conducting the pretest a few weeks before the study, the idea being that participants will not associate the pretest with the study itself. The disadvantage of this solution is that in the time interval between the pretest and the actual treatment and posttest, participants' knowledge may change, making the results unreliable. A modification of this is to have a shorter time differential, but that, of course, weakens the original issue—that of not revealing the topic of the study. A second solution, particularly in the case of assessment of discrete language knowledge, is to ensure that the grammatical/lexical point in question is embedded in a much larger test, thereby reducing the likelihood of participants figuring out the scope of the test. If the participants do not guess the topic from the pretest, the study instruments are more likely to produce a valid characterization of their L2 knowledge.

Instructions/Questions. In addition to guarding against the previously discussed threats to internal validity, one must make sure that the instructions are clear and appropriate to the developmental level of the participants in the study. We cannot rely on responses to questions when it is not clear whether the instructions have been adequately understood. For example, on a university application (filled out by native as well as non-native speakers of English) are the following questions "1. Have you ever been expelled, suspended, disciplined, or placed on probation by any secondary school or college you have attended because of (a) academic dishonesty, (b) financial impropriety, or (c) an offense that harmed or had the potential to harm others?" and "2. Have you ever been convicted of a criminal offense (including in juvenile court) other than a minor traffic violation or are there criminal charges pending against you at this time?" This is followed by "If circumstances arise in the future (until the time you begin attending classes) that make your answers to the above questions inaccurate, misleading, or incomplete, you must provide the Office of Admissions with up-

dated information.” As can be seen, the language is overly sophisticated for those whose English language abilities are not nativelike, and the content (e.g., juvenile court) is inappropriate for a wide range of students who may come from countries with different legal systems. In second language research, the instructions and questions should be appropriate to the level of linguistic and cultural knowledge of those who are taking the test or filling out a questionnaire.

This section has dealt with threats to internal validity. A summary of ways in which such threats can be minimized includes:

- Consider *participant characteristics* that may be relevant to the research questions and elicitation techniques, including but not limited to:
 - Language background.
 - Past language learning experiences.
 - Proficiency level.
 - Specific features and/or skills being examined.
- Consider the issue of *participant mortality*. Make decisions about it before carrying out your research, and justify your solution with respect to:
 - Research design.
 - Research questions.
 - How significant the loss of data would be.
 - (Then be sure to report on this in your research article.)
- Be aware of the possibility that the experimentation itself may affect the results through:
 - Hawthorne and halo effects.
 - Fatigue and boredom of participants.
 - Practice effects of the test material.
- Get the participants’ perspectives after the experiment to ascertain if extra-experimental factors may have impacted their behavior.
- Use a control group to balance *maturational factors* against any long-term requirements of the study.
- Consider how the participants’ performance might be affected by:
 - Physical environment of the study.
 - Characteristics of the researcher.
- Ensure the comparability of pre- and posttests.

- Don't give away the goals of the study.
- Make sure that the instructions are clear and appropriate to the developmental level of the participants.

In the next section, we turn to another type of validity, that known as external validity.

4.6.7. External Validity

All research is conducted within a particular setting and using a specific set of characteristics (e.g., second year L1 English learners of French at X university). However, most quantitative research is concerned with broader implications that go beyond the confines of the research setting and participants. The participants chosen for any study form a research population. With external validity, we are concerned with the generalizability of our findings, or in other words, the extent to which the findings of the study are relevant not only to the research population, but also to the wider population of language learners. It is important to remember that a prerequisite of external validity is internal validity. If a study is not conducted with careful attention to internal validity, it clearly does not make sense to try to generalize the findings to a larger population.

4.6.7.1. Sampling³

The basis of generalizability is the particular sample selected. We want our particular group of participants to be drawn randomly from the population to which we hope to generalize. Thus, in considering generalizability, we need to consider the representativeness of the sample. What this means is that each individual who could be selected for a study has the same chance of being selected as does any other individual. To understand this, we introduce the concept of random sampling.

Random Sampling. Random sampling refers to the selection of participants from the general population that the sample will represent. In most second language studies, the population is the group of all language learners, perhaps in a particular context. Quite clearly, second language researchers do not have access to the entire population (e.g., all learners of

³As mentioned earlier, the sampling procedures discussed in this section relate primarily to quantitative studies. Qualitative research is discussed in chapter 6.

Spanish at U.S. universities), so they have to select an accessible sample that is representative of the entire population.

There are two common types of random sampling: simple random (e.g., putting all names in a hat and drawing from that pool) and stratified random sampling (e.g., random sampling based on categories). Simple random sampling is generally believed to be the best way to obtain a sample that is representative of the population, especially as the sample size gets larger. The key to simple random sampling is ensuring that each and every member of a population has an equal and independent chance of being selected for the research. However, simple random sampling is not used when researchers wish to ensure the representative presence of particular subgroups of the population under study (e.g., male versus female or particular language groups). In that case, stratified random sampling is used.

In stratified random sampling, the proportions of the subgroups in the population are first determined, and then participants are randomly selected from within each stratum according to the established proportions. Stratified random sampling provides precision in terms of the representativeness of the sample and allows preselected characteristics to be used as variables. In some types of second language research it might be necessary, for example, to balance the number of learners from particular L1 backgrounds in experimental groups. For other sorts of second language questions it might be important to include equal numbers of males and females in experimental groups, or to include learners who are roughly equivalent in terms of amount and type of prior instruction or length of residence in the country where the research is being conducted. As an example, assume that one is conducting a study on the acquisition of Arabic passives by speakers of English. Let's further assume that in Arabic language programs, there is a mixture of heritage speakers (those learners who have been exposed to Arabic prior to formal language study through family situations) and nonheritage speakers. Of the students who are available for the study, it turns out that 75% are heritage speakers, making it unlikely that the results will be generalizable to all learners of Arabic. To avoid this problem, the researcher could decide to obtain a sample containing 50% heritage learners and 50% nonheritage learners and randomly select accordingly. This would also make possible what might be an important comparison—that of heritage versus nonheritage learners.

There is yet another approach to sampling, called cluster random sampling. Cluster random sampling is the selection of groups (e.g., intact second language classes) rather than individuals as the objects of study. It is

more effective if larger numbers of clusters are involved. In larger-scale second language research, for example, it might be important to ensure that roughly equal numbers of morning and evening classes receive the same treatments; however, as with any method, the research question should always drive the sampling choice.

How does one obtain a random sample? As mentioned earlier, the principle that should guide selection is that each member of the population has an equal and independent chance of being selected. The purest way of obtaining a true random sample is to take all members of the possible sample, assign each a number, and then use a random number table (available from most statistics books) or a computer-generated random number table (for example, using Microsoft Excel). The following is a small random number table:

068273	241371
255989	213535
652974	357036
801813	313669
188238	987762
858182	324564
539567	010407
874905	076754
705832	752953
394208	866085
532487	980193
717734	499039
965606	256844
442732	809259
128056	843715
398907	972289
999451	782983
016511	525925
980529	329844
657643	501602
123905	385449
941465	573504
311991	088504
594989	631367
163091	221076

If, for example, you have an available population of 99 but you only want to use 35 individuals for your study, you could assign each member a number from 1 to 99 and then use a random number generator to select the first 35. If the first number generated is 77, the person who has been assigned 77 will be part of the data pool. Let's assume that the second number generated is 55. The person who has been assigned 55 will also be a member of the data pool. This continues until 35 numbers have been generated. Alternatively, using the random number table just presented, you could decide to use the last two digits (or the first two or the middle two) and select the first 35 numbers that fall between 01 and 99 until you have the 35 individuals that you need for your study. Starting from the left column and using the last two digits, you would select 73, 89, 74, 13, 38, 82, 67, and so on until you had 35 participants.

Nonrandom Sampling. Nonrandom sampling methods are also common in second language research. Common nonrandom methods include systematic, convenience, and purposive sampling. Systematic sampling is the choice of every n th individual in a population list (where the list should not be ordered systematically). For example, in organizing a new class where learners have seated themselves randomly in small groups (although one must be sure that the seating was truly random rather than in groups of friends/acquaintances), teachers often ask learners to count themselves off as As, Bs, and Cs, putting all the As into one group and so on. In a second language study, researchers could do the same for group assignments, although it would be important that the learners were seated randomly.

Convenience sampling is the selection of individuals who happen to be available for study. For instance, a researcher who wanted to compare the performance of two classes after using different review materials might select the two classes that require the review materials based on the curriculum. The obvious disadvantage to convenience sampling is that it is likely to be biased and should not be taken to be representative of the population. However, samples of convenience are quite common in second language research. For example, researchers may select a time and a place for a study, announce this to a pool of potential participants, and then use those who show up as participants. These learners will show up depending on their motivation to participate and the match between the timetable for the research and their own schedules and other commitments.

In a purposive sample, researchers knowingly select individuals based on their knowledge of the population and in order to elicit data in which they are interested. The sample may or may not be intended to be representa-

tive. For example, teachers may choose to compare two each of their top-, middle-, and lower-scoring students based on their results on a test, or based on how forthcoming these students are when answering questions about classroom processes. Likewise, a researcher may decide to pull out and present in-depth data on particular learners who did and did not develop as a result of some experimental treatment in order to illustrate the different pathways of learners in a study. Some consequences of non-random sampling are discussed later in this chapter.

4.6.7.2. Representativeness and Generalizability

If researchers want the results of a particular study to be generalizable, it is incumbent upon them to make an argument about the representativeness of the sample. Similarly, it is important to describe the setting. A study conducted in a university setting may not be generalizable to a private language school setting. It is often the case that to protect the anonymity of participants, one makes a statement such as the following about the location of the study: “Data were collected from 35 students enrolled in a second-year Japanese class at a large U.S. university.” It is important to minimally include this information so that one can determine generalizability. Private language school students may be different from students at large universities, who may in turn be different from students at other types of institutions.

When choosing a sample, the goal is usually that the sample be of sufficient size to allow for generalization of results, at least for most non-qualitative sorts of research. It is generally accepted that larger samples mean a higher likelihood of only incidental differences between the sample and the population. To reflect this, many statistical tests contain built-in safeguards that help prevent researchers from drawing unwarranted conclusions.

Novice researchers often wonder how many learners are “enough” for each group or for their study overall.⁴ In second language research,

⁴At a large university, a chemist, a physicist, and a statistician were meeting with their Provost in a conference room to explain the real-life applications of their disciplines. During the meeting, a fire broke out in a wastebasket. The physicist whipped out a calculator and began crunching numbers, explaining, “I’m calculating the amount of energy that must be removed in order to stop the combustion.” The chemist thoughtfully examined the fire and jotted down some notes, explaining, “I’m figuring out which reagent can be added to the fire to prevent oxidation.” The Provost seemed impressed at the speed of their reactions and exclaimed, “I had no idea that there could be such immediate real-world applications of your disciplines to a situation like this.” Meanwhile, the statistician pulled out a book of matches and began to set all the other wastebaskets on fire. The shocked Provost demanded, “What are you doing? Are you crazy?” “No, not at all,” replied the statistician, “It’s just that we won’t understand anything until we have a larger N!”

participant numbers vary enormously because of the wide range of different types of research conducted. These research types can range from an intensive experiment including several treatments, pretests, immediate posttests, and multiple delayed posttests, all entailing complex and finely grained linguistic analyses, to a large-scale second language testing study, in which simple numerical before and after scores may be utilized for hundreds of learners. In their text directed at educational research, Fraenkel and Wallen (2003) provided the following minimum sample numbers as a guideline: 100 for descriptive studies, 50 for correlational studies, and 15 to 30 per group in experimental studies depending on how tightly controlled they are. We must remember, however, that research in general education tends to have access to (and to utilize) larger pools than second language research. In second language studies, small groups are sometimes appropriate as long as the techniques for analysis take the numbers into account.

As we have said, a sample must be representative of the population in order for the results to be generalizable. If it is not representative, the findings have limited usefulness. If random sampling is not feasible, there are two possible solutions: First, thoroughly describe the sample studied so that others can judge to whom and in what circumstances the results may be meaningful. Second, as we also discussed in chapter 1, conduct replication studies (and encourage the same of others) wherever possible, using different groups of participants and different situations so that the results, if confirmed, may later be generalized.

4.6.7.3. Collecting Biodata Information

When reporting research, it is important to include sufficient information to allow the reader to determine the extent to which the results of your study are indeed generalizable to a new context. For this reason, the collection of biodata information is an integral part of one's database. The major consideration is how much information to collect and report with respect to the participants themselves. In general, it is recommended that the researcher include enough information for the study to be replicable (American Psychological Association, 2001) and for our purposes in this chapter enough information for readers to determine generalizability. However, the field of second language research lacks clear standards and expectations for the reporting of data, and instances of underreporting are frequent.

In reporting information about participants, the researcher must balance two concerns. The first is the privacy and anonymity of the partici-

Name _____ Research code _____
 Gender: ___ Male ___ Female Age ___ First language(s) _____
 E-mail address _____ Phone number _____
 For how many years have you studied English? _____
 How old were you when you started to study English? _____

Where have you studied English? (tick as many as needed)	How long? (years)	Native English speaker? (yes/no)
_____ Kindergarten	_____	_____
_____ Elementary school	_____	_____
_____ Lower high school	_____	_____
_____ Upper High school	_____	_____
_____ Language schools	_____	_____
_____ Private Tutoring	_____	_____

What English classes are you studying in now? (class numbers and names)

What English classes will you be taking next semester? (Class numbers and names)

Are you studying English anywhere else now? Where? What are you studying (TOEFL, grammar)?

What was your score on the English test of the entrance exam? _____

Have you ever taken the TOEFL test? Yes ___ No ___ What was your score? _____

How many hours per week do you spend using English outside class to ...

Do homework	0	1-2	3-4	5-6
Prepare for quizzes and exams	0	1-2	3-4	5-6
Listen to language tapes	0	1-2	3-4	5-6
Read for fun	0	1-2	3-4	5-6
Listen to music	0	1-2	3-4	5-6
Watch TV, videos & movies	0	1-2	3-4	5-6
Talk to friends	0	1-2	3-4	5-6
Talk to tourists	0	1-2	3-4	5-6
Talk to family members	0	1-2	3-4	5-6

(continued on next page)

Have you ever been to an English-speaking country (UK, Canada, USA, Australia, etc.)?
 Yes ___ No ___

If yes, how long were you there? _____ What did you do there? _____

Have ever been to a country where you spoke English to communicate (Japan, Malaysia, Vietnam, etc.)? Yes ___ No ___

If yes, how long were you there? _____

Besides your first language and English, do you know any other languages? Yes ___ No ___

If yes, which languages? _____

How well do you know them? _____

FIG. 4.1. Sample biodata form.

pants; the second is the need to report sufficient data about the participants to allow future researchers to both evaluate and replicate the study. There are no strict rules or even guidelines about what information should be obtained in the second language field; because of this, exactly what and how much detail is obtained will depend on the research questions and will vary for individual researchers.

It is generally recommended that major demographic characteristics such as gender, age, and race / ethnicity be reported (American Psychological Association, 2001), as well as information relevant to the study itself (e.g., the participants' first languages, previous academic experience, and level of L2 proficiency). Additional information that might be important for a study on second language learning could include the frequency and context of L2 use outside the classroom, amount of travel or experience in countries where the L2 is spoken, learners' self-assessment of their knowledge of the target language, and the participants' familiarity with other languages. Additional information sometimes requested on biodata forms are facts that, although not appropriate for reporting, are necessary for carrying out the research, such as contact information and the association of the participant's name with a code number. *The Publication Manual of the American Psychological Association* also suggested that in reporting information about participants, selection and assignment to treatment groups also be included. The *Manual* further pointed out that "even when a characteristic is not an analytic variable, reporting it may give readers a more complete understanding of the sample and often proves useful in meta-analytic studies that incorporate the article's results" (American Psychological Association, 2001, p. 19).

A sample biodata form appears in Fig. 4.1. As can be seen from the form, depending on the data collection situation some of the questions might require explanations. Not all learners would automatically understand "first

language(s),” for example. Does it mean chronologically first? Does it mean *best* language? They might be more easily able to answer a question about which language they speak at home, or a more specific question about the first language learned and still spoken, or they might understand the term “mother tongue.”

In devising forms for the collection of biographical data, it is important for researchers to balance their need for answers to the questions that could impact their study with requests for extra information that take time to elicit and explain. However, biographical information can be very important when selecting participants; for example, the form in Fig. 4.1 might elicit information about visits to English-speaking countries from even those learners who self-selected into a study on the basis of being beginners, but who then perform at a much higher level than the other learners in the study. This could be important in interpreting results. When selecting the precise questions it is necessary to consider how the data from the *bodata* form will be analyzed. For example, the form in Fig. 4.1 might be useful if one wants to categorize the amount of time spent using the L2 into four categories (none, little, moderate, a lot). A researcher might have used the number of hours on the form to ensure that her category of moderate was the same for all respondents (3–4 hours). However, if one is going to quantify these numbers across participants, these numbers are not easy to work with, particularly if one wants to combine categories into subcategories (e.g., listening). In other words, if a researcher were interested in listening, the categories “listen to language tapes,” “listen to music” and “watch TV” might be combined. The difficulty in interpretation comes when trying to add the numbers. If someone responded 1–2, 3–4 and 3–4, the actual number of hours spent listening could be between 7 and 10, a range likely to be too great to be useful.

There may, however, be instances in which generalizability is not an issue. For example, if one is concerned about making curriculum changes or changes in the way assessment takes place in a particular language program, a research study may be conducted within the borders of that program. The results may turn out to be interesting enough to publish, but it should be understood that the results may or may not be applicable to other contexts and that it is only through empirical study in other contexts that one can determine the generalizability of the original findings.

In this section, we have pointed out that it is often difficult to ensure external validity but have shown ways to minimize threats to external validity. Following is a summary of ways in which one can deal with such threats:

- Random sampling.
- Stratified random selection.
- Systematic, convenience, and purposive sampling.
- Sufficient descriptive information about participants.
- Description of setting.
- Replication of study in a variety of settings.

4.7. RELIABILITY

Reliability in its simplest definition refers to consistency, often meaning instrument consistency. For example, one could ask whether an individual who takes a particular test would get a similar score on two administrations of the same test. If a person takes a written driving test and receives a high score, it would be expected that the individual would also receive a high score if he or she took the same written test again. We could then say the test is reliable. This differs from validity, which measures the extent to which the test is an indication of what it purports to be (in this case, knowledge of the rules of the road). Thus, if someone leaves the licensing bureau having received a high score on the test and runs a red light not knowing that a red light indicates “stop,” we would say that the test is probably not a valid measure of knowledge of the rules of the road. Or, to take another example, if we want to weigh ourselves on scales and with two successive weighings find that there is a 10-pound difference, we would say that the scales are not reliable (although many of us would undoubtedly take the lower weight as the true weight!). In this section, we discuss a number of ways that one can determine rater reliability as well as instrument reliability.

4.7.1. Rater Reliability

The main defining characteristic of rater reliability is that scores by two or more raters or between one rater at Time X and that same rater at Time Y are consistent.

Interrater and Intrarater Reliability. Because these concepts are dealt with in greater detail in chapter 8 on data coding, this section on general reliability provides only a simple introduction. In many instances, test scores are objective and there is little judgment involved. However, it is also common in second language research for researchers to make judgments about data. For example, one might have a dataset from which one wants to extract language

related episodes (LREs), defined as “any part of a dialogue in which students talk about the language they are producing, question their language use, or correct themselves or others” (Swain & Lapkin, 1998, p. 326). We want to make sure that our definition of LREs (or whatever construct we are dealing with) is sufficiently specific to allow any researcher to identify them as such.

Interrater reliability begins with a well-defined construct. It is a measure of whether two or more raters judge the same set of data in the same way. If there is strong reliability, one can then assume with reasonable confidence that raters are judging the same set of data as representing the same phenomenon.

Intrarater reliability is similar, but considers one researcher’s evaluations of data, attempting to ensure that the researcher would judge the data the same way at different times—for example, at Time 1 and at Time 2, or even from the beginning of the data set to the end of the data set. To do this, one essentially uses a test–retest method (see sec. 4.7.2); two sets of ratings are produced by one individual at two times or for different parts of the data. Similar to interrater reliability, if the result is high, then we can be confident in our own consistency (see chap. 8 for a discussion of ways to calculate interrater reliability).

4.7.2. Instrument Reliability

Not only do we have to make sure that our raters are judging what they believe they are judging in a consistent manner, we also need to ensure that our instrument is reliable. In this section, we consider three types of reliability testing: test–retest, equivalence of forms of a test (e.g., pretest and posttest), and internal consistency.

Test–Retest. In a test–retest method of determining reliability, the same test is given to the same group of individuals at two points in time. One must carefully determine the appropriate time interval between test administrations. This is particularly important in second language research given the likelihood that performance on a test at one time can differ from performance on that same test 2 months later, because participants are often in the process of learning (i.e., do not have static knowledge). There is also the possibility of practice effects, and the question of whether such effects impact all participants equally. In order to arrive at a score by which reliability can be established, one determines the correlation coefficient⁵ between the two test administrations.

⁵A correlation coefficient is a decimal (between 0 and 1) that indicates the strength of relationship between two variables. A high correlation coefficient indicates a strong relationship. Correlations are discussed in chapter 9.

Equivalence of Forms. There are times when it is necessary to determine the equivalence of two tests, as, for example, in a pretest and a post-test. Quite clearly, it would be inappropriate to have one version of a test be easier than the other because the results of gains based on treatment would be artificially high or artificially low, as discussed earlier. In this method of determining reliability, two versions of a test are administered to the same individuals and a correlation coefficient is calculated.

Internal Consistency. It is not always possible or feasible to administer tests twice to the same group of individuals (whether the same test or two different versions). Nonetheless, when that is the case, there are statistical methods to determine reliability; split-half, Kuder-Richardson 20 and 21, and Cronbach's α are common ones. We provide a brief description of each.

Split-half procedure is determined by obtaining a correlation coefficient by comparing the performance on half of a test with performance on the other half. This is most frequently done by correlating even-numbered items with odd-numbered items. A statistical adjustment (Spearman-Brown prophecy formula) is generally made to determine the reliability of the test as a whole. If the correlation coefficient is high, it suggests that there is internal consistency to the test.

Kuder-Richardson 20 and 21 are two approaches that are also used. Although Kuder-Richardson 21 requires equal difficulty of the test items, Kuder-Richardson 20 does not. Both are calculated using information consisting of the number of items, the mean, and the standard deviation (see chap. 9). These are best used with large numbers of items.

Cronbach's α is similar to the Kuder-Richardson 20, but is used when the number of possible answers is more than two. Unlike Kuder-Richardson, Cronbach's α can be applied to ordinal data.

4.8. CONCLUSION

In this chapter, we have dealt with some of the general issues that must be considered in designing a research project, such as the importance of properly identifying, operationalizing, and controlling variables, ensuring the internal and external validity of the study, and determining reliability. In the next chapter we deal in greater detail with design.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Read the following brief description of an experiment. Are the conclusions that this researcher came to valid? Why or why not?

The researcher wanted to compare the effectiveness of (a) instruction based on principles of cognitive linguistics coupled with an in-class teacher-led drill, with (b) instruction based on principles of task-based language teaching coupled with independent task work. Both types of instruction focused on the acquisition of locative constructions by second language learners. The experiment was carried out during a 7-week term at the English Language School where the researcher was teaching. Six grammar teachers were assigned to six classes of 15 students each. Two classes were classified as beginning, two as intermediate, and two as advanced according to a placement test that consisted of listening comprehension, reading comprehension, and general grammar. The researcher randomly assigned one teacher at each level to each of the experimental instructional treatments. Students were given an essay to write based on some pictures that produced contexts for locative constructions at the beginning and end of the 7-week term. Each teacher scored his or her students' use of locative constructions based on the number of correct forms over the number of forms attempted. At the end of the 7 weeks, the experimenter collected both sets of scores and compared them. She found that students whose teachers conducted in-class drill sessions had relatively fewer problems with locatives at the end of the session. She therefore concluded that classroom drill is superior to independent task work for the development of correct locative forms in second language learners.

2. In chapter 1, section 1.3.2, we discussed research questions and hypotheses. We listed five of each. Rewrite the hypotheses as null hypotheses.
3. Find three research articles published in different journals, or think of three research articles with which you are familiar. What are the dependent and independent variables in the studies? Is there a moderator or an intervening variable?
4. In the articles you discussed in question 3, what kinds of scales (nominal, ordinal, or interval) were used? Describe the scales and state why they are of the type you identified.

5. The following is a description of a hypothetical study:

A teacher found that, in general, his students (English L1 speakers learning Italian) produced more accurate and more fluent second language speech—in terms of grammatical accuracy and lack of pauses—when describing pictures that had a predictable order than when describing a random assortment of pictures. Following are the characteristics of the participants in this study:

<i>Name</i>	<i>M/F</i>	<i>Age</i>	<i>Accuracy Score</i> (1 is low; 6 is high)		<i>Fluency Score Expressed</i> <i>in Total Pause Length</i>	
			<i>Predictable</i>	<i>Random</i>	<i>Predictable</i>	<i>Random</i>
Miranda	F	21	4.4	3.8	4.6	4.8
Octavia	F	22	5.2	4.6	3.2	3.5
David	M	26	4.5	4.7	3.9	3.4
Josh	M	28	4.1	4.6	3.8	3.2
Aaron	M	31	4.7	4.9	4.4	3.9
Seth	M	27	4.6	4.8	4.8	4.5
Ethan	M	25	4.5	4.4	5.2	4.6
Rebecca	F	18	3.8	3.2	1.0	1.0
Stefania	F	17	3.9	3.4	1.0	1.2
Kerry	F	24	4.4	3.9	4.1	4.6
Rabia	F	24	4.4	3.9	1.9	2.0
John	M	32	1.9	2.0	4.1	4.6
Rachel	F	23	2.4	1.6	4.7	4.9
Thomas	M	19	1.0	1.2	2.4	1.6
Natasha	F	24	4.8	4.5	4.5	4.4
Marc	M	23	1.0	1.0	4.4	3.9
Ella	F	20	5.0	4.1	4.1	4.6
Michael	M	19	3.2	3.5	5.0	4.1
Robert	M	20	4.1	4.6	4.0	3.8
Sharona	F	22	3.9	3.2	4.5	4.7

- What are the dependent, independent, and moderator variables? Do you think that the moderator variable is important in this study? Why or why not? To determine this you might want to consider the averages for male versus female participants on the different conditions.
6. Find a study in which the researcher clearly operationalizes a variable. What needed to be operationalized? How was it operationalized? Are there alternative ways it could have been operationalized? What are they?
 7. Suggest ways to operationalize the following constructs:
 - a. Amount of input.
 - b. Attentiveness in class.
 - c. Interest in class.
 - d. Language proficiency.
 - e. Prior language learning experience in L2.
 - f. Prior language learning experience in other languages.
 8. Are the following directional or nondirectional hypotheses?
 - a. German learners of Russian will outperform Spanish learners of Russian in the acquisition of case marking.
 - b. There will be more instances of relative clause production by Hindi learners of French than by Japanese learners of French.
 - c. There will be a difference between Hindi learners of French and Japanese learners of French in the amount of relative clause production.
 - d. Students in a lab setting will show more evidence of circumlocution than those in a classroom setting.
 - e. There will be a difference in learning between students who are presented with models of language as opposed to those who are given recasts.
 - f. Students who are induced to make an error through overgeneralization followed by immediate correction will outperform those who are presented with a correct form from the outset.
 9. Based on a topic that interests you, write three directional and three non-directional hypotheses.
 10. To which does each of the following refer: nominal scale, ordinal scale, interval scale, or ratio scale?
 - a. A scale in which equal differences are truly equal differences of the variable being measured.

- b. A scale that can't measure quantitative variables.
 - c. A scale with a true zero point.
 - d. A scale that indicates relative rankings.
11. Consider the data in the accompanying table. The data are from two test administrations in which English students' knowledge of French relative clauses was being tested. The researcher was pilot-testing the instrument to see if it was reliable. Testing was done on two successive days so that it would be unlikely that any learning took place. On each test there was a maximum of 20 points (2 points for each of 10 sentences). Partially correct responses were awarded 1 point.

	<i>Test 1</i>	<i>Test 2</i>
Sally	18	12
Marie	15	12
Jean	10	14
Howard	15	16
Janice	14	14
Robert	19	18
Drew	8	15
Andrew	11	7
Marc	6	12
Grace	10	10

Given these scores, are you confident that you have developed a reliable test? Why or why not?

12. Read the following abstract and answer the questions.⁶
- Article title:* Evidence in Favor of a Broad Framework for Pronunciation Instruction
- Article source:* *Language Learning*, 1998, 48, 393–410.
- Authors:* Tracy Derwing, Murray Munro, and Grace Wiebe
- Abstract:* We had native English-speaking (native speaker) listeners evaluate the effects of 3 types of instruction (segmental accuracy; general speaking habits

⁶This problem was provided by Charlene Polio (adapted by Mackey & Gass).

and prosodic factors; and no specific pronunciation instruction) on the speech of 3 groups of English as a second language (ESL) learners. We recorded their sentences and extemporaneously produced narratives at the beginning and end of a 12-week course of instruction. In a blind rating task, 48 native English listeners judged randomized sentences for accentedness and comprehensibility. Six experienced ESL teachers evaluated narratives for accent, comprehensibility, and fluency. Although both groups instructed in pronunciation showed significant improvement in comprehensibility and accentedness on the sentences, only the global group showed improvement in comprehensibility and fluency in the narratives. We argue that the focus of instruction and the attentional demands on speakers and listeners account for these findings.

Segmental = phonetic features, i.e., vowel and consonant sounds, and no prosodic features

Prosodic features = generally stress, intonation, pitch, volume, i.e., suprasegmental features

Extemporaneously produced narrative = without any planning

Blind rating task = The evaluators did not know which set of sentences or narratives belonged to which treatment group.

Questions:

- a. What is the independent variable in this study?
- b. The dependent variable, pronunciation (accentedness, comprehensibility, fluency), was measured in many different ways in this study. Do you think that those measures were categorical, ordinal, or interval? Explain.
- c. How was the issue of validity in assessing the dependent variable dealt with in this study?
- d. How would you check the reliability of the measures of pronunciation?
- e. Was the blind rating done to ensure the internal or external validity of the study? Explain.

13. Provide brief descriptions of second language studies in which you might want to use the following kinds of sampling—random, stratified random, cluster random, systematic, convenience, and purposive—and explain why.
14. Why does (or why doesn't) replicating research make that research more generalizable?
15. Consider a study of the relationship between peer responses to L2 writing and linguistic accuracy, and explain why a researcher might want to obtain biographical data on the following:
 - Age.
 - First language.
 - Length of residence.
 - Amount and type of prior L2 writing instruction.

What else could a researcher find out about learners' profiles to inform this study?

Designing a Quantitative Study

This chapter deals with design types in quantitative research. We begin by introducing the materials that are used together with ways of placing individuals into groups. We then move to the central part of the chapter when we focus on ways of designing a study, including pretest/posttest designs, posttest-only designs, time-series designs, and one-shot designs. Throughout, we discuss the considerations researchers should make when designing a study for a given topic and population.

5.1. INTRODUCTION

In previous chapters, we discussed the need to state clear and answerable research questions and provided information on the construction of questions and the selection of variables. In this and the following chapter, we deal with the design of a study. The focus of this chapter is quantitative research; in chapter 6 we focus on qualitative research (and briefly discuss studies that combine both approaches and present quantitative and qualitative results).

Quantitative research can be conceptually divided into two types: associational and experimental. What is common in both types is that researchers are attempting to determine a relationship between or within variables. The goal of associational research is to determine whether a relationship exists between variables and, if so, the strength of that relationship. This is often tested statistically through correlations, which allow a researcher to determine how closely two variables (e.g., motivation and language ability) are related in a given population. Associational research is not concerned with causation, only with co-occurrence. In experimental studies, researchers deliberately manipulate one or more variables (independent variables) to determine the effect on another variable (dependent variable). This manipulation is usually described as a treatment and the researcher's goal is to determine

whether there is a causal relationship. Many types of experimental research involve a comparison of pretreatment and posttreatment performance.

In this chapter, we describe some of the issues that need to be considered in both associational and experimental research. It is important to note from the outset that all research designs involve decisions at each step of the way and typically, many of these decisions are fraught with compromises: If I do X, then I cannot do Y. Thus, in designing a research project, there is often a cost/benefit analysis to undertake.

Example:

You want to carry out research on the acquisition of past tense forms following recasts in an online chat session. One of your research questions involves the effects of recasts on participants of different language backgrounds (the languages vary in the degree to which their past tense system is similar to the target language past tense system). You need participants at a level of proficiency that is not too high, so that there is some room for learning. Also, your participants' proficiency levels cannot be too low, because otherwise they will not be able to carry out the task.

As you begin to look for participants of the right language background and the right proficiency level, you realize that this is more difficult than you had originally imagined. If you have students of an appropriate proficiency level, the right native languages are not available. Participants of the appropriate native languages are not available at the proficiency level that is right for your study.

Thus, you will be forced to make a compromise and possibly eliminate one of the variables of your study.

5.2. RESEARCH MATERIALS

One of the key components of designs involves the actual materials used, and it is here that we begin. In chapter 3 we presented various data elicitation techniques and discussed ways to avoid some of the pitfalls that occur in data elicitation, the most important of which is to ensure that the data obtained can truly address the research questions. Likewise, all materials need to be pilot-tested, as discussed in chapter 1, in order to ensure that what you want to elicit is in fact what you are eliciting. Following is a list of some of the ways that materials can be the source of a problem:

- Insufficient tokens.

Example 1 (Spanish copular ser and estar): You want to determine whether learners of Spanish understand the difference between the two copular forms (there are subtle semantic and pragmatic differences between the two) and only one example of each is elicited by a 10-item sentence completion task. It would be difficult to draw reliable conclusions from these data because there are not enough examples.

Example 2 (English past tense): You try to elicit examples of the past tense in English as a Second Language by using a narrative task. You give instructions to participants to recount a past event, but they *describe* rather than *narrate*, again providing you with data with very few examples of past tense. For example, when asked to describe his favorite birthday party, one learner begins by narrating, “My brother gave me a party” and then continues by describing, “My brother is a very good brother. He always do many things for me. He always call me, he always visit with me. He give me very good party.” Such descriptions do not constitute past tense contexts.¹

- Task appropriateness for the elicitation of target structure.

Example (Italian noun–adjective agreement): You want to elicit noun–adjective agreement in Italian as a foreign language, but in your task there are very few opportunities to describe items, or learners can easily avoid the structure, as in the following exchange:

NNS: *Quella cosa (f) è blu (f. & m.)*

“That thing is blue”

NS: *Quale cosa?*

“What thing”

¹A further difficulty with this situation is that we cannot be certain that these examples do not constitute actual attempts at the past. Given the possibility that the learner could produce the past tense, but contexts for the past to occur are not adequately targeted by the prompt (or that there is ambiguity), this prompt should probably be eliminated in favor of one for which there is greater certainty of eliciting the target structure if the learner is at the correct level to produce it.

NNS: *Libro* (m)
 “book”

If the target structure can be easily avoided in the task, it is inappropriate for use in the study.

- Imprecise instructions.

Example (English relative clause formation): The following data elicitation exercise was designed to produce relative clauses.

- *Version 1:* Combine the two sentences below making one sentence. Start with the first sentence.

The boy loves chocolate. I saw the boy.

Expected response:

The boy that I saw loves chocolate.

Actual response: *The boy loves chocolate and I saw the boy.*

- *Version 2:* Combine the two sentences below making one sentence. Do not use the words *and*, *or*, or *but*. Start with the first sentence.

The boy was running. I saw the boy.

Expected response:

The boy that I saw was running.

Actual response:

The boy was running even though I saw him.

- *Version 3:* Combine the two sentences below making one sentence. Do not use the words *and*, *or*, *but*, *although*, *even though*, or *however*. Start with the first sentence.

Result: These instructions worked well except that the participants did not always begin with the first sentence. Thus, the targeted structure was not elicited. (Fortunately, this “violation” of the instructions—see Gass, 1980—turned out to have interesting implications for relative clause acquisition.)

- Insufficient examples for learners to understand what to do.

Example (English questions): A spot-the-difference task was used to study English as a Second Language question use among leaner dyads.

Instructions: You and your partner each have a picture. Your pictures are very similar, but there are 10 differences. Don't show your partner your picture; ask questions to find the differences.

NNS1: What the man is doing on the floor?

NNS2: The man is sleeping.

NNS1: Is not floor but is next beach, is it, next to beach.

NNS2: I don't have in my picture.

NNS1: Is the same picture.

NNS2: No is not similar.

(Adams, personal communication, 2004)

In this situation, NNS2 did not understand that she and her partner had different pictures. Rather than looking for the differences, she became upset when her partner did not confirm her description of her picture. Although tasks and directions may seem obvious to researchers, they may be new and confusing for each participant. Sufficient examples should always be provided.

The preceding examples illustrate some problems that can occur. As mentioned earlier, the best and perhaps the only way to ensure that your materials will allow you to answer the research questions that you have posed is to pilot-test them. In your research report, you can then justify your choice of research materials by discussing the pilot-test data. An example of this can be seen in the selection of vocabulary items in a study of attention (Gass, Svetics, & Lemelin, 2003). The authors selected vocabulary items and pilot-tested them on a separate group of learners to make sure that the words would be unfamiliar to a group of learners similar to those in the experiment. They stated in their article: "Five words were selected for focus in this study that pilot-testing showed were unlikely to be known by the participants" (p. 512). It would not have been possible to ask participants in the actual study if they knew the words or in this particular experiment to test their knowledge of the words in some way, because the fact of asking or testing could have served as an impetus for learning.

5.3. INTACT CLASSES

In chapter 4, we discussed the ways in which randomization can enhance the experimental validity of a study. However, there are situations when

randomization of individuals may not be feasible. For example, in second language research we often need to use intact classes for our studies, and in these cases the participants cannot be randomly assigned to one of the experimental or control groups. Intact classes are commonly and often by necessity used in research for the sake of convenience. Consider the following design (Gass et al., 1999), which used intact classes to examine the effects of task repetition on learners' subsequent production:

Research questions:

- Does task repetition yield more sophisticated language use?
- Will more accurate and/or sophisticated language use carry over to a new context?

Method: Show film clips a different number of times to different Spanish classes (at the same proficiency level) followed by the showing of a new film clip of the same genre.

Groups:

- *Experimental Group 1:* This class saw one episode three times followed by a fourth viewing of a different episode.
- *Experimental Group 2:* This class also had four viewings of a film, but each video was different (the first was the same as Group 1's first and the fourth was the same as Group 1's fourth).
- *Control group:* This class saw only two episodes: the first and the fourth.

In this case, the alternative to using intact classes would have been to randomly assign individuals to one of the three groups (2 experimental and 1 control). Given that this study took place on a university campus where students have varied academic (and sometimes work) schedules, it would have been unlikely that all the individuals assigned to Group 1 would have been able to meet at one time to view the video. The alternative would have been multiple video showings. This would represent a very significant time burden for the researchers in addition to the logistical problems of scheduling. To put this into a time frame, consider a classic experimental/control group study, with 20 learners in each group and treatments that last 1 hour each. If two intact classes participated (one designated control and the other experimental), data elicitation would require 2 hours. If participants from both classes were randomly assigned to control and experimental groups and the research had to conduct individual treatment sessions, data

elicitation would require 40 hours. This obviously represents a greater strain on human resources.

The use of intact classes, although not typical of experimental research, may have the advantage of enhancing the face validity of certain types of classroom research. For example, if the effects of a particular instructional method are investigated, an existing classroom may be the most ecologically sound setting for the research (for more in-depth discussion of classroom research, see chap. 7).

If intact classes are used, the researcher should carefully consider how the classes are assigned to treatment groups. One way of dealing with nonrandomization of individuals is to use a semi-randomization procedure by arbitrarily assigning classes to one treatment or another. However, there are other considerations as well. Suppose that you want to compare five sections of German as a foreign language classes. Unless students are randomly placed in sections, it might be the case that there is a different profile for students who opt to take an 8:00 A.M. class when compared with those who opt to take a 5:00 P.M. class. One group may include more learners with off-campus jobs, for example, whereas another group may include those who are exclusively studying full-time.

5.4. COUNTERBALANCING

Counterbalancing refers to an experimental design in which the ordering of test items or tasks are different for different participants. To look at this more closely, consider the following two designs, in which the researcher wants to investigate the effect of writing topics (the independent variable) on the amount of coherence that is produced in the form of transition words (the dependent variable). For reasons of logistics, in this fictional study we cannot randomly assign individuals and we cannot do a pretest to determine group equivalence.

Design 1: Nonrandom Assignment of Participants to Groups

Research question: Do different L2 writing topics yield different amounts of transitional words?

Design:

- Group 1: Compare your two best teachers.
- Group 2: Describe a traumatic event.

- Group 3: Argue in favor of language programs in elementary schools.

Analysis: Compare the number of instances of transitional words among the three groups.

In this design, participants were not assigned randomly and, because there was no pretest, some might question the extent to which the groups are comparable (see further discussion on non-pretest designs in section 5.5.3.2). One way to compensate for this lack of comparability is to counterbalance by having all groups do all tasks in a different order, as in Design 2. Because each individual does all tasks, and the order is different for each group, the issue of the possible lack of comparability due to ordering effects can be minimized.

Design 2: Counterbalanced Design

Research question: Do different writing topics yield different amounts of transitional words?

Design:

- Group 1: (1) Compare your two best teachers; (2) describe a traumatic event; (3) argue in favor of language programs in elementary schools.
- Group 2: (1) Describe a traumatic event; (2) compare your two best teachers; (3) argue in favor of language programs in elementary schools.
- Group 3: (1) Argue in favor of language programs in elementary schools; (2) compare your two best teachers; (3) describe a traumatic event.
- Group 4: (1) Compare your two best teachers; (2) argue in favor of language programs in elementary schools; (3) describe a traumatic event.
- Group 5: (1) Describe a traumatic event; (2) argue in favor of language programs in elementary schools; (3) compare your two best teachers.
- Group 6: (1) Argue in favor of language programs in elementary schools; (2) describe a traumatic event; (3) compare your two best teachers.

Analysis: Compare the number of instances of transitional words in each of the tasks.

In the analysis, the data are averaged across groups. In other words, all comparison and contrast results would be averaged across all six groups so that there is no ordering effect. The disadvantage is that one must rely on the goodwill and attentiveness of the participants to participate in three tasks, especially if all the data are being collected at one time. Another disadvantage is that because of the toll (e.g., in fatigue and boredom) that the experiment takes on participants (assuming each task takes approximately 10–15 minutes), one must limit the number of topics and thus one cannot get a full range of topics. In addition to counterbalancing treatments, researchers can also explicitly test for order effects, particularly when there are relatively few treatments, as in this example.

5.5. RESEARCH DESIGN TYPES

5.5.1. Correlational (Associational) Research

Correlation can be used in different ways: for example, to test a relationship between or among variables, and to make predictions. Predictions are dependent on the outcome of a strong relationship between or among variables. That is, if variables are strongly related, we can often predict the likelihood of the presence of one from the presence of the other(s). Correlation is often used in survey-based research (see chap. 3), although it is by no means limited to that research area. Following are examples of two types of survey-based correlational research (one relational and one predictive), both from a large survey-based study of motivation by Dörnyei and Clément (2001):

Research question: Are student motivational characteristics related to language choice?

Context:

1. Motivational characteristics (e.g., direct contact with L2 speakers, cultural interest, integrativeness, linguistic self-confidence) were collected through questionnaires from more than 4700 Hungarian students.
2. Information was gathered on their language of choice in school (e.g. American vs. British English, German, French, Italian, Russian).

- Analysis:* The study was set up so that the relationship between these variables could be examined.
- Research question:* Was integrativeness (represented by questions such as “How important do you think learning these languages is in order to learn more about the culture and art of its speakers?” “How much do you like these languages?” and “How much would you like to become similar to the people who speak these languages?”) a predictor of language choice?
- Analysis:* The follow-up analysis showed that integrativeness was the best predictor of language choice.

The specific statistical analyses will be discussed in chapter 9.

5.5.2. Experimental and Quasi-Experimental Research

In chapter 4, we introduced the concept of random assignment of participants and the need to ensure that each participant in a particular population has an equal and independent opportunity for selection. Randomization is usually viewed as one of the hallmarks of experimental research. Design types can range from truly experimental (with random assignment) to what is known as quasi-experimental (without random assignment). Clearly, some design types are more prototypical of one end of the range than the other. In this section we deal with both types, beginning with those that include random assignment of individuals.

A typical experimental study usually uses comparison or control groups to investigate research questions. Many second language research studies involve a comparison between two or more groups. This is known as a between-groups design. This comparison can be made in one of two ways: two or more groups with different treatments; or two or more groups, one of which, the control group, receives no treatment.

5.5.2.1. Comparison Group Design

In a comparison group design, participants are randomly assigned to one of the groups, with treatment (the independent variable) differing between or among the groups.

Example: A researcher wants to investigate whether aural input or input through conversational interaction yields better L2 learning.

- Group 1: Hears a text with input containing the target structure.
- Group 2: Interacts with someone who provides input on the target structure.

Assuming a pretest/posttest design (see later discussion), the results of the two groups would be compared, with inferences being made as to the more appropriate method of providing information to learners. In comparison research, more treatment groups can be added to the study if the research question is elaborated. The following example suggests a slightly different research question with a more elaborate design:

Example: A researcher wants to investigate to what extent aural input, input through conversational interaction, or a combination of aural and conversational input yields better L2 learning.

- Group 1: Listens to a text containing the target structure.
- Group 2: Interacts with someone who provides input on the target structure.
- Group 3: Receives some input through listening and some through interaction.

Were this researcher to add yet another question, the design could have a fourth group:

Example: A researcher wants to investigate to what extent aural input, input through conversational interaction, a combination with aural input followed by interaction, or a combination with interaction followed by aural input yields better learning.

- Group 1: Hears a text with input containing the target structure.
- Group 2: Interacts with someone who provides input on the target structure.
- Group 3: Receives some input first through listening and then through interaction.
- Group 4: Receives some input first through interaction and then through listening.

5.5.2.2. Control Group Design

The second standard type of experimental design is a control group design. This is similar to the comparison group design, with the important difference that one group does not receive the treatment. The control group would typically take the same pretest and posttest as would the experimental groups, but would not have the same treatment in between tests. For control groups, some researchers may want to provide some language activity or input (of course, different from the treatment) in which the participants are doing something else with language. This is to ensure that it was the treatment, not the mere fact of doing something, that led to any change.

One aspect that all researchers grapple with in second language research is how to control for outside exposure to the language. This is much easier to control in a foreign language environment than in a second language environment. In a foreign language setting, control for exposure can often be accomplished simply by ensuring that the particular language focus is not covered in the syllabus during the time of treatment. Another way to prevent external input influencing the results of the study is not to have long periods of time between testing sessions (although there are instances when long periods of time are desirable, as in delayed posttests when testing longer-term effects of treatment), or, minimally to be able to argue that if there is additional exposure, the groups are equivalent in this regard. Each researcher must be cognizant of the problem and must determine how to deal with it in an appropriate manner.

In sum, a true experimental design will have some form of comparison between groups. The groups will differ in terms of some manipulation of the independent variable to examine the effect of manipulation on the dependent variable. Assignment will be random, or as random as possible, to avoid threats to internal validity caused by participant characteristics.

5.5.3. Measuring the Effect of Treatment

There are a number of ways of designing a study using the design characteristics discussed earlier. We focus on two: pretest / posttest (with and without delayed posttests), and posttest only.

5.5.3.1. Pretest/Posttest Design

In many second language studies, participants are given a pretest to ensure comparability of the participant groups prior to their treatment, and a

posttest to measure the effects of treatment. In chapter 3, we discussed the need to ensure that our measures are assessing what we intend them to assess. Once we have determined that our measures are indeed appropriate for our research question, there is a further question to be addressed: Is the pretest comparable in difficulty to the posttest? If the pretest turned out to be more difficult than the posttest, participants might demonstrate artificially greater improvement; if the pretest turned out to be easier, participants might demonstrate artificially less improvement. We discussed the importance of comparability of tests in chapter 4, where we outlined one solution in the form of a study that randomly assigned some sentences to the pretest and others to the posttest, with the sentence assignment differing for each learner. In this way, the threat of test bias is reduced. Another way of accomplishing this is to test all sentences on a comparable group of individuals to ensure that no test items are more difficult than others and, if they are, to place a comparable number of difficult ones and a comparable number of easy ones in each test.

In a pretest/posttest design, researchers can determine the immediate effect of treatment. However, clearly, the real question for studies of second language learning is to address to what extent a treatment truly resulted in learning. From what we know about learning, it is a process that may begin with a particular treatment, but it is not always clear that the effects of that learning are long-lasting. To measure the longer-term effects, researchers often want to include delayed posttests in addition to the immediate posttests. In fact, this is becoming increasingly common in second language research. With delayed posttests, a test comparable to the posttest (and pretest, if there is one) is administered one or more times at predetermined times after the treatment. Often this is 1 week following the first posttest and then 2 weeks later and even 2 or 3 months later. The advantage of delayed posttests is that one gets a wider snapshot of treatment effects; the disadvantages are that there is a greater likelihood of losing participants, extra-experimental exposure will be greater, and there will be maturation. It is important to make an a priori principled decision as to how many posttests a participant must take before eliminating data.

5.5.3.2. Posttest-Only Design

In some instances, it is undesirable to give a pretest because the pretest itself might alert participants to what the treatment concerns. In most cases, we do not want our participants to know or guess the purposes of the treatment (see chap. 4 for a discussion of the problems of giving the goals of the study away and ways to address this). It is important to recognize that there

are limitations to this type of design, the main one being that a researcher cannot be sure if there is initial comparability of groups. This becomes an issue particularly when measuring improvement as a result of the manipulation of an independent variable. In posttest-only designs the focus of study is usually performance and not development. Following are suggestions for ways to address the issue of lack of initial comparability measures, so that one can make assumptions about group comparability:

- When using intact classes (see sec. 5.3), if at all possible (which may be difficult if one is using more than two or three intact classes), select classes that meet at roughly the same time (8:00 A.M. classes versus 4:00 P.M. classes).
- Match learners on other measures that one can gain from a background questionnaire, for example:
 - Gender.
 - Age.
 - Years of language study.
 - Language(s) spoken in the home.
 - Class grades.
 - Teachers' ratings.
 - Placement test.
- Match learners on a variable that you can argue is related to the dependent variable. For example, if you can argue that attitude and motivation are related and your study is about motivation, you could match participants on the variable of attitude.
- Match learners on performance on a first treatment.
- Base the investigation on language features that have not previously appeared in the syllabus or in the textbook (relevant for foreign language environments).
- Use low-frequency words for vocabulary studies.

5.5.4. Repeated Measures Design

As mentioned earlier in our discussion of randomization, a common way of dealing with the problem of nonrandomization is through a repeated measures design, in which all tasks or all treatments are given to different individuals in different orders. The basic characteristic of a repeated measures design (also known as a within-group design) is that multiple measurements come from each participant. An example of a repeated measures design fol-

lows. In this example (from Gass, 1994), data on the target structure are elicited from all learners at two different points in time. This is similar to the example of a counterbalanced design to elicit information on the effect of writing topic on the use of transitional words that was presented earlier. In that study, all learners produced writings on each of the different topics.

- Purpose:* To assess the reliability of using acceptability judgments in L2 research.
- Research question:* Do learners make similar judgments of acceptability at two different points in time?
- Method:* L2 learners of English judged relative clauses on a 7-point scale (including a 0 point). Sentences were randomized for each participant. The judgment exercise was repeated at a 1-week interval with sentences of the same grammatical structure.
- Scoring:* Each person was given a score from -3 (definitely incorrect) to +3 (definitely correct) for each sentence at each time period. The responses at Time 1 and Time 2 were compared to determine the extent to which responses were similar / different at Time 1 and Time 2.

In this repeated measures study, each participant's score at Time 1 was compared with his or her score at Time 2. The research question itself dictated a repeated measures design.

5.5.5. Factorial Design

A factorial design involves more than one independent variable and can occur with or without randomization. A factorial design allows researchers to consider more than one independent variable, generally moderator variables (see chap. 4 for a discussion of variables). The example that follows shows a possible factorial design for the investigation of topic effect on word count/sentence.

Research questions:

- Question 1: Do different writing topics yield different word counts?
- Question 2: Does L1 background yield different word counts?
- Question 3: Do writing topics and L1 background interact to yield different word counts?

Makeup of group:

- Native Korean speakers: 19.
- Native Spanish speakers: 22.
- Native Arabic speakers: 21.

Analysis for Question 1: Compare results across all topics for all learners (total number of learners is 62). This is a main effect.

Analysis for Question 2: Compare results across L1 background groups for all writing topics. This is a main effect. The three groups would consist of: 19 (Korean), 22 (Spanish), 21 (Arabic).

Analysis for Question 3: Compare the groups to see if the pattern is the same for all three language groups. This is an interaction effect.

A results table of this factorial design presented might look like Table 5.1.

These group means are different, as demonstrated by Fig. 5.1. Without statistical testing, of course, it is not possible to determine whether these differences are significant. Fig. 5.1 is a possible graph of the results of a factorial design. Figure 5.1 provides a graphic representation of the data and suggests that the type of writing task influences the average words per sentence for each group. It also suggests that the effect of the type of writing task is similar for each group.

5.5.6. Time-Series Design

Time-series designs are frequently used with small groups of learners. A time-series design involves repeated observations (both pretest and posttest)

TABLE 5.1
Average Number of Words per Sentence, by Topic and Group

Topic	Native Language			Total
	Korean	Spanish	Arabic	
Compare two favorite teachers	6	12	5	23
Describe a traumatic experience	12	18	10	40
Argue in favor of language programs	10	15	8	33
Total	28	45	23	96

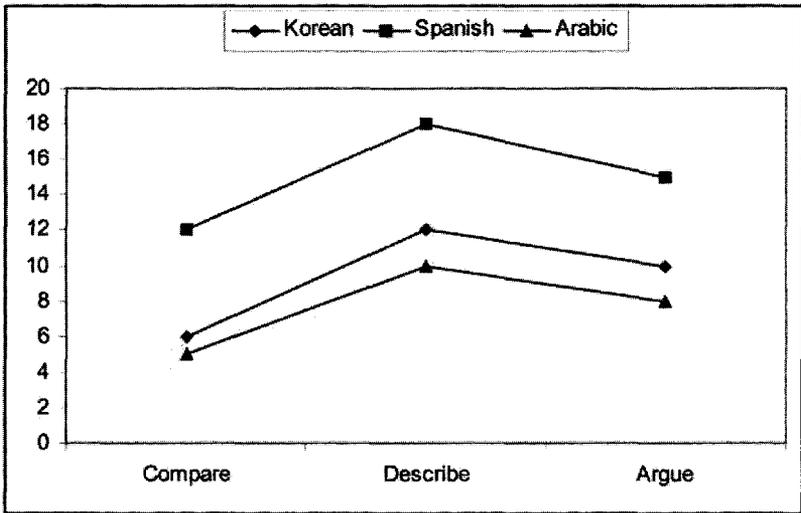


FIG. 5.1. Words per sentence, by topic and L1 background.

over a set period of time. Before the treatments, a set of observations is made to establish a baseline. Following the treatment, further observations are made to ascertain the effects of the treatment. A typical pattern for a time-series design is the following, where O refers to observations:

$$O_1 O_2 O_3 O_4 O_5 \text{ Treatment } O_6 O_7 O_8 O_9 O_{10}$$

If, for example, a researcher finds comparable results for all the observations prior to the treatment, there is evidence to determine the normal patterns that a particular group exhibits before the treatment. Similarly, one can obtain a sense of the patterns that occur following the treatment. An example cited by Mellow, Reeder, and Forster (1996) employed data from Kennedy (1988). Kennedy was interested in the effect of dyadic grouping on discourse. The participant pool consisted of four 4-year-old children and the study took place over a 10-week period. These deliberately paired dyadic interactions in which mathematics was discussed were compared with interactions in free play. Treatment was introduced at different points for each of the four dyads. Fig. 5.2 (from Mellow et al., 1996) shows the results of one of the measures (number of NNS turns per episode) and provides information on the nature of the design. As can be seen, the preintervention time for each participant differed, ranging from 3 weeks (#1) to 6 weeks (#4). Similarly, the treatment

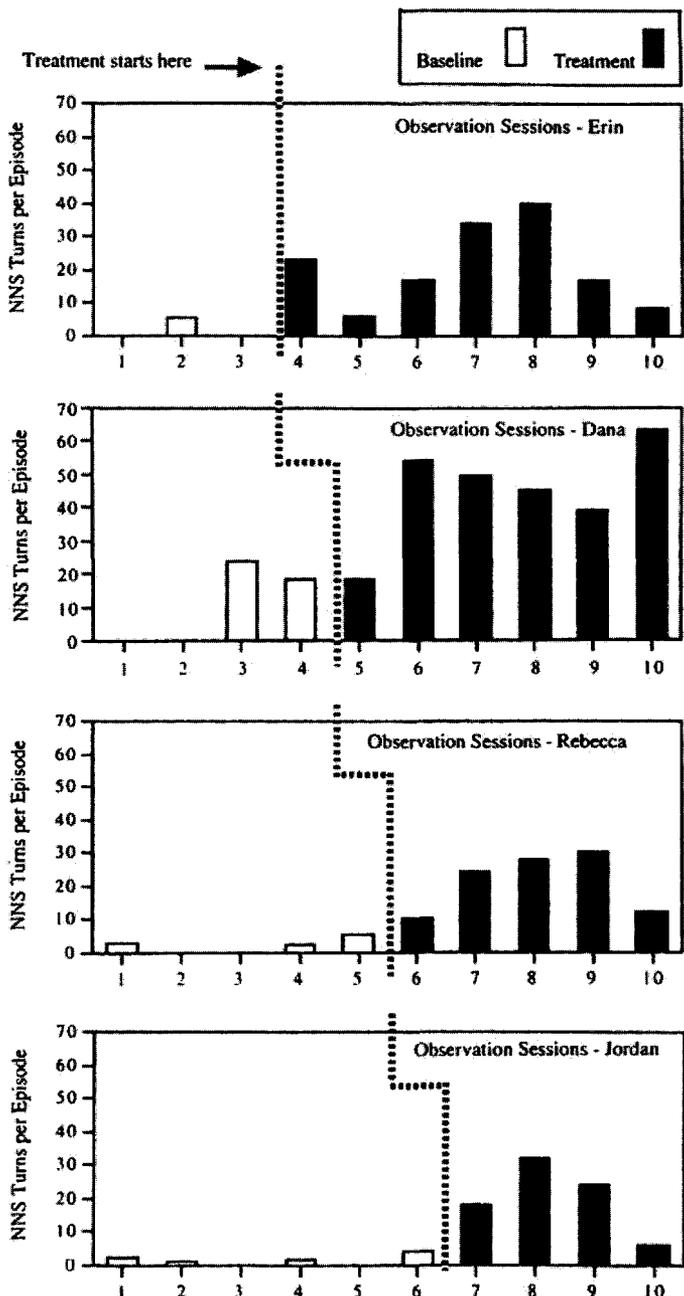


FIG. 5.2. Number of NNS turns per episode. Source: Mellow, J. D., Reeder, K., & Forster, F. (1996). Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition*, 18, 340. Copyright © 1996 by Cambridge University Press. Reprinted with the permission of Cambridge University Press.

time differed for each participant. Using a design such as this, one can look at each treatment session and determine the rank for whatever is being measured (the dependent variable). For example, if one looks at Session 4, one can see that the highest score comes from the only participant who had already received the treatment (#1).

In summary, a time-series design can overcome some of the problems typical in second language research, in which there can be both small numbers and noncomparability of individuals at the outset. It also reduces some of the problems inherent in research that does not utilize a control group. In a time-series design, as can be seen from Fig. 5.2, multiple samples can be taken from an individual/groups before and after the treatment, allowing a researcher to generate certain baseline information about each participant/groups, thereby allowing a comparison among individuals/groups at different points of the pretest–treatment–posttest continuum. For example, we can see that at Time 4, Erin (the only one with treatment at this time) had more turns per episode than the other three. In addition, the consistency in samples for each individual before treatment increases the confidence in the effects of the treatment and hence increases the internal validity.

Mellow et al. (1996) listed various options for variations in time-series designs, of which we mention two:

- One can gather baseline data on an individual in a classroom followed by treatment and then withdrawal of treatment. If the learner reverts to pretreatment behavior following withdrawal, one would have confidence in the effects of the treatment.
- With multiple individuals, each individual can receive treatment at a different time.

Time-series designs have a great deal of flexibility. Mellow et al. (1996) isolated four reasons for utilizing this design type:

- It is practical—it can be used even with small numbers of participants.
- It reduces Hawthorne effect (see chaps. 4 & 7) because many instances of data collection are used.
- It can be used as a means of exploration and hypothesis-generation.
- Given longitudinal design and many instances of measurement, it provides a richer picture of development.

In general, using a time-series design does not eliminate the possibility of maturational effects, but the use of multiple subjects, treated at different times, may make findings more convincing.

5.5.7. One-Shot Designs

We have limited our discussion to the most common design types used in second language research, but there are others as well. One of these is what we refer to as a “one-shot” design. This is not usually considered part of a true experimental paradigm, because there is no treatment. Nevertheless, one-shot designs are often used in second language research within the UG or processing paradigms (see chap. 3 for discussions of each of these) when the study does not have a pretest/posttest design, but simply raises questions along the lines of: What do learners know at this particular point in time, or how do learners interpret sentences in an L2? An example of this type of research comes from White (1985), who was following up on a well-motivated theoretical question from formal linguistics.²

Research question: Is there a relationship among the following three types of sentences produced by Spanish speaking learners of English?

That-trace:

<i>Spanish</i>	<i>English</i>
<i>Quien dijeste que vino?</i>	<i>*Who did you say that came?</i>
who you said that came	<i>Who did you say came?</i>
“Who did you say came?”	

Subject-verb inversion:

<i>Spanish</i>	<i>English</i>
<i>Vino Juan.</i>	<i>John came.</i>
came Juan	<i>*Came John.</i>
“Juan came.”	

²This study involved a comparison between Spanish and French native speakers in which the French group served as a control, but the point is that a one-shot design can be used when one is interested in considering what a group of learners knows about the TL at one point in time.

Null Subjects

<i>Spanish</i>	<i>English</i>
<i>Anda muy ocupada.</i>	<i>*Is very busy.</i>
is very busy	<i>She is very busy.</i>
'She is very busy.'	

Method: Present participants with sentences (both grammatical and ungrammatical) and ask for acceptability judgments.

Results: Spanish speakers continue to cluster these sentences in their L2, particularly at lower levels of proficiency.

Questions other than UG-based questions lend themselves to this design type as well. Consider the research question from Gass, Mackey, and Ross-Feldman (2003).

Research question: Is there a difference between the amount of interaction in classroom contexts versus lab contexts?

Operationalization of interaction: Language-related episodes, recasts, negotiation.

Method: Three tasks given to participants in their regular classrooms. Identical three tasks given to a different and comparable group of participants in a laboratory context.

Results: No difference in the amount of interaction in the two contexts.

Thus, not all studies require a design that uses a control group and an experimental group or that necessitates a pretest and a posttest. Following is a summary of the different research design types discussed in this chapter.

Group assignment

- Comparison group: Compares two or more types of treatment.
- Control group: Compares treatment group(s) to a group that receives no treatment (on targeted structures).

Measurement of treatment effects

- Pretest/posttest design: Carry out pretreatment testing to ensure comparability of groups, and/or to establish learners' initial level of ability/knowledge, posttreatment testing to measure learning.
- Posttest design: Posttreatment testing to measure differences among groups or to measure learners' ability/knowledge after treatment.
- Repeated measures design: Elicit multiple samples from the same learners over time.
- Factorial design: Examine the effects of multiple variables on a dependent variable.
- Time-series design: Make multiple observations before treatment to establish initial baseline and after treatment to measure effects of treatment.
- One-shot designs: Characterize learner knowledge or behavior at one particular time.

5. 6.FINALIZING YOUR PROJECT

As has become clear in this chapter, there are many areas to consider when designing a research project. Following is a checklist of some of the most common areas to think about as you are putting together your design (but of course, as always, pilot, pilot, pilot):

- Are your groups matched for proficiency?
- If you are using a particular type of task (e.g., listening), are your groups matched for (listening) abilities?
- Are your participants randomized?
- If intact classes are used, are their treatments randomly assigned?
- Are your variables clear and well described?
- Do you have a control group?
- Are control groups and experimental groups matched for everything but the specific treatment (including the time spent on the control and experimental tasks)?
- Have you described your control and experimental groups?
- Do you have a pretest?
- If you are testing development, do you have a posttest or even multiple posttests?

- If using a repeated-measures design, are the treatments counterbalanced?

5.7. CONCLUSION

This chapter has dealt with some basic issues of design in quantitative research. We have discussed the most commonly used design types in experimental and quasi-experimental research, identifying them along such dimensions as participant assignment to treatment group, testing, and variables included. In the next chapter, we focus on issues of methodology in qualitative research.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Which of the following are researchable questions through experimental research (E)? Which are library research questions (questions that are better researched not through original data, but through searching prior sources; L)? Which are research questions that would require partial or complete redefinition before the question could be investigated (R)?
 - a. Why should the government finance English as a Second Language classes for refugee families?
 - b. What are the characteristics of a good language learner?
 - c. Does articulatory explanation improve learners' ability to produce the /i/ versus /I/ distinction in English?
 - d. Do high-anxiety students make fewer errors on compositions than do low-anxiety students?
 - e. Do students remember more pairs of antonyms than pairs of synonyms when one member of the pair is presented in the first language and the other is presented in the second language?
 - f. Does the use of a bilingual dictionary help foreign language students learn more vocabulary than does use of a monolingual dictionary in the L2?
 - g. Does a student's perception of similarities/differences between his or her first language and the second language influence transfer of syntactic forms from the first language to the second language?

- h. Is extended listening with delayed oral practice more effective than is a total skills approach in initial language learning?
2. You are completing a study on the effects of participation in a volunteer aiding program on later performance in ESL practice teaching. What is the dependent variable? What is the independent variable? You also believe that type of class in which aiding was done (elementary school, adult school, university class) might have some relationship to success in student teaching. Identify this variable type.
3. You have asked each of your ESL students to go out in the “real world” and make five complaints during the next week. They will judge their success on a 5-point success-to-failure scale. During the previous week, half of these students had watched a videotape of an American woman returning a watch to a store, complaining that it did not work properly. You want to know if the model helped with self-rated success. What is the dependent variable in this study? The independent variable? All your students are adult women from a variety of language backgrounds. Some of them work in factories, some in stores, and some in offices. How would you identify these variables? You decide that this might be a good pilot project on ESL learner success in speech events. Suggest another variable that might be important in such a study. How would you measure this variable?
4. Consider the following conclusions. Are they valid? Why or why not? If not, what would make them more convincing?
 - a. Second language learners who identify with the target culture will master the language more quickly than will those who do not. (*Evidence 1*: a case study of an unsuccessful language learner who did not identify with the target language; *Evidence 2*: five case studies of unsuccessful language learners who did not identify with the target language and five case studies of successful language learners who did identify with the target language; *Evidence 3*: same as #2, but the data are accompanied by verbal reports from learners showing that this is indeed an important connection.)
 - b. Immigrants are more law abiding than native-born citizens. (*Evidence*: an analysis of court records.)
 - c. Affective relationships between teacher and students influence proficiency gains. (*Evidence*: a longitudinal ethnographic study of an inner-city high school class.)

- d. Input followed by interaction promotes better learning than does interaction followed by input. (*Evidence*: two groups of 50 each—the group with input followed by interaction outperformed the group with interaction followed by input on an immediate posttest and two subsequent posttests.)
5. Read the following abstract and answer the questions.³

Article title: “The Influence of Task Structure and Processing Conditions on Narrative Retellings”

Article source: *Language Learning*, 1999, 49, 93–120.

Authors: Peter Skehan and Pauline Foster

This article explores the effects of inherent task structure and processing load on performance on a narrative retelling task. Task performance is analyzed in terms of competition among fluency, complexity, and accuracy. In a study based on 47 young adult low-intermediate subjects, the fluency of performance was found to be strongly affected by degree of inherent task structure; more structured tasks generated more fluent language. In contrast, complexity of language was influenced by processing load. Accuracy of performance seemed dependent on an interaction between the two factors of task structure and processing load. We discuss which aspects of performance receive attention by the language learner. The implications of such cross-sectional results for longer-term language development are considered.

Questions:

- a. What is the research question addressed in this study?
- b. Is this an experimental study? Why or why not?
- c. What are the independent variables in this study?
- d. For each independent variable, state what kind of variable it is—that is, what kind of scale is being used? Justify your response.
- e. What are the dependent variables?
- f. How might one operationalize each of the dependent variables?

³This problem was provided by Charlene Polio (adapted by Mackey & Gass).

Qualitative Research

The importance and utility of qualitative methods is increasingly being recognized in the field of second language research. This chapter begins with a discussion of the nature of qualitative research and how it differs from other approaches. Next, commonly used methods for gathering qualitative data are outlined, including case studies, ethnographies, interviews, observational techniques, verbal protocols, and diaries/journals. We conclude the chapter with a discussion of practical considerations in conducting and analyzing qualitative research, including issues such as credibility, transferability, dependability, triangulation, and quantification.

6.1. DEFINING QUALITATIVE RESEARCH

The term *qualitative research* is associated with a range of different methods, perspectives, and approaches. As Mason (1996) pointed out, “qualitative research—whatever it might be—certainly does not represent a unified set of techniques or philosophies, and indeed has grown out of a wide range of intellectual and disciplinary traditions” (p. 3). Nonetheless, for the purposes of this chapter, we attempt to present a general definition of qualitative research in the second language field and to outline several of its key characteristics. Briefly defined, the term *qualitative research* can be taken to refer to research that is based on descriptive data that does not make (regular) use of statistical procedures. Detailed definitions of qualitative research usually include the following characteristics:

- *Rich description*: The aims of qualitative researchers often involve the provision of careful and detailed descriptions as opposed to the quantification of data through measurements, frequencies, scores, and ratings.

- *Natural and holistic representation*: Qualitative researchers aim to study individuals and events in their natural settings (Tetnowski & Damico, 2001). That is, rather than attempting to control contextual factors (extraneous variables) through the use of laboratories or other artificial environments, qualitative researchers tend to be more interested in presenting a natural and holistic picture of the phenomena being studied. This picture includes both the broader sociocultural context (e.g., the ideological orientations of the speech community as a whole) as well as micro-level phenomena (e.g., interaction within the classroom).
- *Few participants*: Rather than using a large group of (generally randomly selected) participants with the goal of generalizing to a larger population like quantitative researchers, qualitative researchers tend to work more intensively with fewer participants, and are less concerned about issues of generalizability.
- *Emic perspectives*: Qualitative researchers aim to interpret phenomena in terms of the meanings people attach to them—that is, to adopt an emic perspective, or the use of categories that are meaningful to members of the speech community under study. For instance, it might be inappropriate in some cultures for students to laugh at, question, or to make eye contact with their teachers. A qualitative researcher would aim to take this into account when investigating student affect in the classroom. Emic perspectives can be distinguished from the use of etic (or outsiders') categories and frameworks to interpret behavior. Etic perspectives are more common in quantitative studies.
- *Cyclical and open-ended processes*: Qualitative research is often process-oriented, or open ended, with categories that emerge. The research often follows an inductive path that begins with few perceived notions, followed by a gradual fine-tuning and narrowing of focus. In contrast, quantitative research usually begins with a carefully defined research question that guides the process of data collection and analysis. Thus, whereas quantitative researchers set out to test specific hypotheses, qualitative researchers tend to approach the research context with the purpose of observing whatever may be present there, and letting further questions emerge from the context.
- *Possible ideological orientations*: Whereas most quantitative researchers consider impartiality to be a goal of their research, some qualita-

tive researchers may consciously take ideological positions. This sort of research is sometimes described as ‘critical,’ meaning that the research may have particular social or political goals. For example, Scollon (2001) argued that critical discourse analysis, a form of qualitative research, is “a program of social analysis that critically analyzes discourse—that is to say language in use—as a means of addressing social change” (p. 139).

- Research questions tend to be general and open ended, and hypotheses may be generated as an outcome of qualitative research rather than in the initial stages of the study. According to Brown (2003), “One of the great strengths often cited for qualitative research is its potential for forming new hypotheses.” (p. 485).

Table 6.1, from Chaudron (2000), provides a useful overview of the distinctions between qualitative and quantitative approaches (also discussed in chap. 1).

Despite the fact that distinctions can be drawn between qualitative and quantitative research as shown in Table 6.1, these two research types are by no means as dichotomous as they sometimes appear to be. Compounding this confusion, it is increasingly common for researchers to present and discuss both quantitative and qualitative data in the same report, or to use methods associated with both types of research in a process sometimes known as split methods or multiple methods.

For example, Sullivan and Pratt (1996) used both quantitative and qualitative methodologies to investigate the effects of computer technology on ESL student writers. The researchers used quantitative approaches to compare student essays in two types of writing environments (classes using computer technology and traditional oral classrooms); qualitative analyses were then used to compare the types and patterns of discourse in those environments. By combining these approaches, Sullivan and Pratt were able to present a more detailed picture of how the computer technology affected the quality of the students’ writing, their patterns of discourse, and their perspectives on the value of the technology.

The growing practice of utilizing qualitative and quantitative data illustrates the fact that these two research approaches should not be viewed as opposing poles in a dichotomy, but rather as complementary means of investigating the complex phenomena at work in second language acquisition.

TABLE 6.1
Distinctions Between Qualitative and Quantitative Approaches

	<i>Qualitative Methods (Ethnography)</i>	<i>Quantitative Methods</i>
Observation & Collection of Data	In data collection, ethnographic research (as the most typical and concrete example of qualitative research) doesn't usually use "instruments," rather "processes" that are supposedly free of bias and prior assumptions: free, prolonged observation, at times "participant observation," open-ended interviews, "triangulation" of information and interpretation, "informant checking," access to existing documents.	The observations in quantitative research (whether tests, attitudes scales of the subjects observed, behaviors categorized and counted according to instruments, etc.) usually are based on an observation scheme or descriptive categories that have been developed prior to the research. Moreover, these observations are made in a planned way, according to an order determined by the design of the research, and with categories that cannot be changed once the research is underway.
Nature of Data	Ethnographic research considers those data most relevant which arise from the natural events in the research context. The topics of greatest interest for qualitative researchers are human behaviors and socio-cultural patterns and norms which underlie the behaviors. Data are viewed in a "holistic" fashion, without attempting to separate them into their components, and preferably following the interpretations of the people who are the object of the research ("emic" interpretations).	Data tend to be limited by the type of observation that is planned, and according to the method of observation; depending on the design and the effects of a treatment, the data usually indicate stability or variability and development in events, attitudes, abilities, skills, knowledge, performance or production, etc., with respect to a language and its use. These are interpreted according to the theoretical model or hypotheses of the researcher, and not necessarily according to the views of the subjects involved ("etic" interpretation).

(continued on next page)

TABLE 6.1 (continued)

	<i>Qualitative Methods (Ethnography)</i>	<i>Quantitative Methods</i>
Use and Development of Theory	The qualitative researcher does not want to verify or prove theories; what she/he attempts is to observe without bias nor narrow perspectives. However, the researcher always takes account of the relevant theories regarding the context or topic under study, and normally will remain aware of her/his own assumptions during observation and interpretation. Proper methodology will include the appropriate degree of "objectivity." In the end, the researcher will develop a "grounded" theory which helps to relate the observations to one another and to larger contexts, or she/he will attempt to revise and perfect the conceptual framework which was adopted at an earlier stage. In the most radical form of qualitative research (from the tradition of phenomenology), causal explanations are not sought, but only a better "understanding" of the phenomenon.	The researcher constructs a design to prove some aspect of a theoretical framework (forming hypotheses about the goals of the research), and the results tend to either confirm or disconfirm the hypotheses. Although it is recognized that the researcher's subjectivity can influence interpretations, in order not to generalize beyond the research context, the design, which includes the means of sampling the subjects, should control the limits of conclusions to be drawn. Thus, a theoretical framework is slowly developed.

Source: Chaudron, C. (2000). Contrasting approaches to classroom research: Qualitative and quantitative analysis of language and learning. *Second Language Studies*, 19(1), 7. Copyright © 2000 by Craig Chaudron. Reprinted with the permission of Craig Chaudron and the Working Papers in Second Language Studies.

6.2. GATHERING QUALITATIVE DATA

As noted earlier, a wide variety of different techniques are used in the collection of qualitative data. As with all methods, the advantages and disadvantages of each technique should be taken into consideration when deciding how to address a specific research question. Here we present an overview of some of the most commonly used qualitative data collection methods, including:

- Ethnographies
- Interviews
- Diaries/journals
- Case studies
- Observational techniques

As discussed earlier, because there is little general agreement in the field about what constitutes qualitative research, some of the data collection techniques we discuss in this chapter are associated with more “descriptive” than truly “qualitative” methods by some researchers. For example, Brown (2003) categorized interviews and questionnaires as part of survey-based research, a distinct category from qualitative and quantitative research, which he referred to as “interpretive and statistical methods.” Also, given that some of the data collection methods described here are associated with particular contexts or overlap with each other, some are also described in other chapters in this text. For example, diaries and journals are also discussed in chapter 7 on second language classroom research contexts, with examples from teachers and learners. Each approach and method can be seen as contributing its own piece of the puzzle in qualitative researchers’ attempts to obtain rich, detailed, participant-oriented pictures of the phenomena under study.

6.2.1. Ethnographies

Although there has been much recent debate concerning the nature of ethnography, it can be said from a second language research perspective that ethnographic research aims “to describe and interpret the cultural behavior, including communicative behavior, of a group” (Johnson, 1992, p. 134) as well as “to give an emically oriented description of the cultural practices of individuals” (Ramanathan & Atkinson, 1999, p. 49), or, in other words, to

carry out research from the participants' point of view, using categories relevant to a particular group and cultural system.

This focus on group behavior and the cultural patterns underlying that behavior is one of the key principles of ethnography identified by Watson-Gegeo in her (1988) review article. Another important principle of ethnographic research is the holistic approach taken to describing and explaining a particular pattern in relation to a whole system of patterns. Hence, ethnography can be viewed as a qualitative research method that generally focuses on the group rather than on the individual, stresses the importance of situating the study within the larger sociocultural context, and strives to present an emic perspective of the phenomena under investigation, with the categories and codes for analysis being derived from the data themselves rather than being imposed from the outside.

Ethnographic approaches have been used in a very broad range of second language research contexts, ranging from ethnographies of schools and language programs to personal accounts and narratives or life histories of learning and teaching (e.g., Duff, 2002; Pavlenko & Lantolf, 2000), home-school discontinuities among Native American children (e.g., Macias, 1987; Philips, 1972, 1983), bilingual language use outside educational settings (Goldstein, 1997), cultural and ideological processes in first and second language learning (King, 2000), and research on specific aspects of the L2 process, such as second language writing in different cultural contexts (e.g., Atkinson & Ramanathan, 1995; Carson & Nelson, 1996). However, as Johnson (1992) noted, one of the main uses of ethnographic research in the second language context "has been to inform us about the ways that students' cultural experiences in home and community compare with the culture of the schools, universities, and communities where they study, and the implications of these differences for second language and culture learning" (p. 135).

A well-known piece of research of this nature, an ethnography of first language communication, was carried out by Shirley Brice Heath (1983), who spent a decade living in what she described as two working class communities in the Carolinas; in her terms, one black community, Trackton, and one white community, Roadville. Using data from these working-class communities, as well as middle-class school-oriented black and white families in the town, Heath's research focused on how the children in her study learned to use language, how their uses of language were related to their literacy, and how their use of language in the home interacted with how they used print. She showed how the different ways of

learning language interacted with the children's integration into academic life. The working-class communities held different expectations and different usage patterns, as well as different attitudes to the mainstream families and schools in terms of uses of language, including what Heath referred to as "ritualized uses of language," such as the assignment of labels to objects, response to questions whose answers were already known to the questioner (often known as display questions), and recitation of discrete points of factual material separated from context. In summary, the language and literacy practices of middle-class families mirrored expectations in the school. These differences in home language and literacy practices had implications for academic success.

The goal of such research is to be emic, detailed, holistic, and situated in context with a focus on exploring how complex factors interact. Ethnographies can profitably make use of methods specifically designed to tap into participants' perspectives and, as such, they often involve or are overlapped with the use of observations, interviews, diaries, and other means of data collection that are discussed in more detail later in this chapter. They also generally involve triangulation of data, which is also discussed later in this chapter.

6.2.1.1. Advantages

One advantage of using an ethnographic approach is that the research questions employed in these studies can be dynamic, subject to constant revision, and refined as the research continues to uncover new knowledge. For example, an ethnographer studying second language writing classrooms may enter the research process with the aim of describing the patterns of interaction between teachers and students and illustrating how those patterns are related to the writing process. However, over the course of many classroom observations, analyses of student essays, and interviews with both students and teachers, the researcher may alter the focus of the study and begin concentrating on the types of feedback that are provided by both teachers and students. Ethnographic approaches are particularly valuable when not enough is known about the context or situation to establish narrowly defined questions or develop formal hypotheses. For example, why does a particular group of heritage language learners do poorly when learning in formally instructed foreign language classroom settings? An ethnographic approach to this question could examine the context, the attitude of the teacher and the students, the influence of home and social

groups, and so on in an attempt to uncover information relevant to addressing the question. If the researcher shares the heritage language background, participant observation could be used; that is, the researcher might be able to participate in the language classes or share social occasions in which the language is used in some way. Because ethnographies typically employ multiple methods for gathering data, such as participant observations and open-ended interviews as well as written products, ethnographic research may be able to provide an holistic, culturally grounded, and emic perspective of the phenomena under investigation.

6.2.1.2. Caveats

In embarking on an ethnographic study, researchers need to be aware of some potential challenges and sensitive issues. First of all, ethnographies involve intensive research over an extended period of time. They require a commitment to long-term data collection, detailed and continuous record keeping, and repeated and careful analysis of data obtained from multiple sources. It is also important for the researcher to realize that ethnographic approaches to research may create potential conflicts between the researcher's roles as an observer and a participant. If the researcher participates in an event he or she is observing, this may leave little time for the carefully detailed field notes that ethnographies may require. This can be rectified to a certain extent by audio and video tape recording. However, and more seriously, the researcher's participation may change the nature of the event (see also the discussion on observations in sec. 6.2.4). Researchers thus need to be aware of how they can supplement and triangulate ethnographic data obtained through participant observation, and they must carefully consider how their dual roles might influence the data collected.

In addition to these practical concerns, there are theoretical issues that the researcher should take into consideration. First of all, it has been argued that an ethnographer's focus on describing a culture is problematic, because "there is no such thing as a social group that is not constantly destabilized by both outside influences and personal idiosyncrasy and agency" (Ramanathan & Atkinson, 1999, p. 45). In its strong form, this criticism implies that any attempt to describe a group is to some extent misguided on the part of the ethnographer. A second theoretical concern about ethnographies concerns the act of writing up the research. Because research reports adhere to certain (culturally influenced) standards of writing, the otherwise accurate picture an ethnographer has recorded may come out skewed. In other words, the

very act of transcribing the events that were observed may inevitably entail a misrepresentation of them. Finally, it is often difficult to generalize the findings of ethnographic research to other problems or settings because of the highly specific nature of such work.

6.2.2. Case Studies

Like ethnographies, case studies generally aim to provide a holistic description of language learning or use within a specific population and setting. However, whereas ethnographies focus on cultural patterns within groups, case studies tend to provide detailed descriptions of specific learners (or sometimes classes) within their learning setting. Case studies are also usually associated with a longitudinal approach, in which observations of the phenomena under investigation are made at periodic intervals for an extended period of time.

Case studies have been used in a wide variety of second language research studies. One well-known longitudinal case study investigating the development of L2 communicative competence is Schmidt's (1983) study of Wes, an ESL learner. Wes was a 33-year-old native speaker of Japanese who had little formal instruction in English. Schmidt studied Wes' language development over a 3-year period when he was residing in Japan but visited Hawaii, the research site, regularly on business. The study focused on a small number of grammatical features, including plural *s*, third-person singular *s*, and regular past tense. Schmidt transcribed conversations between Wes and friends and also transcribed monologues that he asked Wes to produce while at home in Japan. Although Wes attained relatively high levels of pragmatic ability and acculturation (e.g., in the use of formulae such as "So, what's new?" and "Whaddya know?"), he had very limited improvement in terms of linguistic accuracy for the grammatical forms over the 3 years of the study, thus providing evidence for the separability of linguistic and pragmatic competence.

Another well-known case study is Ellis' (1984) investigation of two child learners in an L2 context. J was a 10-year-old Portuguese boy, and R was an 11-year-old Pakistani boy. Both children were learning ESL in London in a program that catered to new arrivals with the aim of preparing them for transfer to high schools. Ellis investigated the learning patterns of the two children in an instructed context, as opposed to the naturalistic context in which Schmidt had studied Wes' language development. In order to examine the learners' use of requests, Ellis visited the classrooms regularly, writing

down the samples of requests the two children produced. Ellis' analysis documented different stages in the children's use of requests over time, also noting their tendency to use formulae. In another second language case study in which the focus was the role of collaborative interaction in second language development, Ohta (1995) examined the language of two intermediate learners of Japanese as a foreign language, finding that their patterns of interaction facilitated a form of scaffolding, or assisted help, that benefited both learners. Obviously, cases can also be individual classes or schools.

6.2.2.1. Advantages

One main advantage of case studies is that they allow the researcher to focus on the individual in a way that is rarely possible in group research. As Johnson (1993) noted, "[I] too often, because of the nature of correlational, survey, and experimental research, and their privileged status in L2 research, very little is learned about individual language learners, teachers, or classes. Case studies stand in sharp contrast to these approaches by providing insights into the complexities of particular cases in their particular contexts" (p. 7). In addition, case studies can be conducted with more than one individual learner or more than one existing group of learners for the purpose of comparing and contrasting their behaviors within their particular context. Case studies clearly have the potential for rich contextualization that can shed light on the complexities of the second language learning process.

6.2.2.2. Caveats

An essential point to bear in mind with case studies, however, is that the researcher must be careful about the generalizations drawn from the study. Although this is true for all forms of research, it is particularly pertinent to case studies, which often employ only a few participants who are not randomly chosen. For this reason, any generalizations from the individual or small group (or classroom) to the larger population of second language learners must be made tentatively and with extreme caution. From a single case study, it may be difficult to recognize idiosyncrasies as such, with the potential that they are misinterpreted as typical language learning behavior. To address this concern, the findings from multiple longitudinal case studies can be combined to help researchers draw firmer conclusions from their research. For example, Wray (2001) summarized 14 case studies that focused on the role of formulaic sequences in child second language acqui-

sition. The cases involved 21 children (12 girls and 9 boys), aged approximately 2 to 10. Based on these multiple case studies, Wray argued that children express themselves holistically in a second language by employing formulaic sequences. In short, case studies may provide valuable insights into certain aspects of second language learning, but single case studies are not easily generalizable.

6.2.3. Interviews

A number of different interview types can be employed to gather data for qualitative research. As noted in the introduction to this chapter, interviews are often associated with survey-based research, as well as being a technique used by many qualitative researchers. In structured (also known as standardized) interviews, researchers usually ask an identical set of questions of all respondents. Structured interviews resemble verbal questionnaires and allow researchers to compare answers from different participants. Less rigid are semistructured interviews, in which the researcher uses a written list of questions as a guide, while still having the freedom to digress and probe for more information. In unstructured interviews, on the other hand, no list of questions is used. Instead, interviewers develop and adapt their own questions, helping respondents to open up and express themselves in their own terms and at their own speed. Unstructured interviews are more similar to natural conversations, and the outcomes are not limited by the researcher's preconceived ideas about the area of interest. Some interviews can also be based around a stimulus—for example, a completed questionnaire or a videotape of a lesson. Focus-group sessions are related to such interviews, and usually involve several participants in a group discussion, often with a facilitator whose goal it is to keep the group discussion targeted on specific topics, again often using a stimulus for discussion, such as a videotape or previously elicited data.

6.2.3.1. Advantages

Interviews can allow researchers to investigate phenomena that are not directly observable, such as learners' self-reported perceptions or attitudes. Also, because interviews are interactive, researchers can elicit additional data if initial answers are vague, incomplete, off-topic, or not specific enough. Another advantage of interviews is that they can be used to elicit data from learners who are not comfortable in other modes. For example,

some learners are more at ease speaking than writing and are more likely to provide extended answers in a conversational format. Depending on the research question and the resources available, interviews can also be conducted in the learner's L1, thus removing concerns about the proficiency of the learner impacting the quality and quantity of the data provided.

6.2.3.2. Caveats

Researchers must also take note of the potential drawbacks of interviews. For example, Hall and Rist (1999) made the point that interviews may involve “selective recall, self-delusion, perceptual distortions, memory loss from the respondent, and subjectivity in the researcher’s recording and interpreting of the data” (pp. 297–298). Multiple interviews—that is, interviewing the same subject more than once, or interviewing many different subjects—is one potential means of addressing such issues. Another concern is that good interviewing is a skill. It may not be easy for novice researchers to conduct unstructured interviews without practice and/or training in drawing participants out, encouraging them to express themselves, and gathering valuable data on the area of interest. Given that participants’ attitudes toward other people can impact what they say, there is also the danger of the so-called halo effect (discussed earlier in chap. 4). This refers to what happens when interviewees pick up cues from the researcher related to what they think the researcher wants them to say, thus potentially influencing their responses. In addition to these concerns, the possibility of cross-cultural pragmatic failure exists. Some questions may be considered inappropriate in particular cultures, and because of the different connotations words carry in different linguistic and cultural contexts, miscommunications may arise.

To address some of these concerns, the following suggestions may be useful in interviewing:

- Be sensitive to (and/or match the interviewer’s characteristics with) the age, gender, and cultural background of the interviewee.
- Encourage open-ended discussion—for example, by keeping silent, or by saying “Anything else?” rather than accepting a first answer as the interviewee’s final and complete response to a question.
- Develop skills in anticipating and addressing communication problems.
- Try to make the interviewee as comfortable as possible. This can be done by conducting the interview in a familiar place, beginning with

- small talk to relax the interviewee, and/or using the L1 if a communication problem arises or if the interviewee so prefers.
- Place the key questions in the middle of the interview, because the interviewee may be nervous in the beginning and tired by the end.
 - Mirror the interviewee's responses by repeating them neutrally to provide an opportunity for reflection and further input.

6.2.4. Observations

As Mason (1996) noted, observation usually refers to “methods of generating data which involve the researcher immersing [him or herself] in a research setting, and systematically observing dimensions of that setting, interactions, relationships, actions, events, and so on, within it” (p. 60). When collecting data using observational techniques, researchers aim to provide careful descriptions of learners' activities without unduly influencing the events in which the learners are engaged. The data are often collected through some combination of field notes (which can involve detailed impressions of the researcher's intuitions, impressions, and even questions as they emerge) and audio or visual recordings (which allow the researcher to analyze language use in greater depth later and to involve outside researchers in the consideration of the data). In second language research, observations have been used in a wide variety of studies, ranging from naturalistic studies to the rather more common classroom observations that are discussed at length in chapter 7.

Different types of observations can be identified, depending on their degree of structure. In highly structured observations, the researcher often utilizes a detailed checklist or rating scale. In a complex L2 environment such as the language school, workplace, or community, a structured observation can facilitate the recording of details such as when, where, and how often certain types of phenomena occur, allowing the researcher to compare behaviors across research contexts in a principled manner. In less structured observations, the researcher may rely on field notes for detailed descriptions of the phenomena being observed, or transcripts of tapes of those events.

6.2.4.1. Advantages

Observations are useful in that they provide the researcher with the opportunity to collect large amounts of rich data on the participants' behavior

and actions within a particular context. Over time and repeated observations, the researcher can gain a deeper and more multilayered understanding of the participants and their context.

6.2.4.2. Caveats

Observations typically do not allow the researcher access to the participants' motivation for their behaviors and actions. For this reason, observations may be most useful when combined with one or more of the other methods discussed in this book (see sec. 6.3.2 on triangulation). However, perhaps the most serious concern is the "observer's paradox" (Labov, 1972). This refers to the fact that although the goal of most observational research is to collect data as unobtrusively as possible, the presence of an observer can influence the linguistic behavior of those being observed. There is also some possibility of the Hawthorne effect (discussed in chaps. 4 and 7), which may occur when learners perform better due to positive feelings at being included in a study. Simply put, if learners realize that they are under observation, their performances might improve because of the fact of that observation.

To minimize these threats, researchers should consider the ways in which they may influence an L2 setting and take steps to mitigate the effect of their presence. For example, if the goal of a study involves observing the use of a second language among immigrants in their workplace, researchers may try to blend into the background of the workplace to make the participants more accustomed to their presence. Another less obtrusive option is participant observation, by which researchers are members of the group they are observing. They play a dual role of observing while fully participating in activities with other group members. Although participant observation can limit the effects of the observer's paradox, it can also be difficult to both observe and participate, as discussed previously. Participant observation is generally most feasible in adult learning contexts where the researcher can easily blend in—for example, in conversation or language exchange clubs. Ethical issues related to participant observations also need to be considered. It is important to keep issues related to partial and full disclosure of the goals of a study in mind, as discussed in chapter 2.

6.2.5. Diaries/Journals

Because learners' reports about their internal processes and thoughts can be elicited by carefully tailoring the questions that researchers choose to ask, ver-

bal protocols and other introspective methods are often used to gather data for qualitative studies (see chap. 3 for additional discussion of verbal protocols). Second language diaries, also referred to as L2 journals or learner autobiographies, can also be used to allow learners, language professionals (and teachers, as discussed in chap. 7 on classroom research) to write about their language learning experiences without the constraints imposed by specific questions.

One well-known diary study in the second language research field was described in Schmidt and Frota's (1986) research on the language learning diary of one of the researchers' experiences learning Portuguese in Brazil. Schmidt used a diary to record his language learning experiences in classes of Portuguese as a second language, as well as in his daily interactions while living in Brazil. His diary included the specific second language forms that he was taught in class, in addition to those he observed during conversations and those that he found himself using. Additionally, he met with his co-researcher, a native speaker of Portuguese, for periodic communicative second language testing. Through an examination of the diary entries and results of the testing, he was able to detect an interesting pattern: Schmidt consistently consciously noticed forms in the input shortly before acquiring them.

Another often-cited study in the second language research field is the work of Schumann and Schumann (1977), who reported on diaries that they kept as they attempted to learn Arabic in North Africa at the introductory level, and as they learned Persian (Farsi) in a U.S. university setting as well as in Iran as intermediate level learners. Schumann and Schumann remarked on the diary itself as "a possible vehicle for facilitating the language learning process" (p. 241). They also pointed out that the detailed records of emotional issues, such as transition anxiety, found in their diaries suggest that individual variables can promote or inhibit second language learning. On the topic of transition anxiety and second language learning, one of the diary entries revealed, "I found one reasonably effective way to control this stress during travel to the foreign country. Enroute to Tunisia and during the first week or so after arrival, I devoted every free minute to working through an elementary Arabic reader Learning more of the language gave me a sense of satisfaction and accomplishment that went a long way toward counter-acting the anxiety" (p. 246). Other diary studies in second language research involve researchers analyzing the diaries of language learners, often in instructed settings.

6.2.5.1. Advantages

In many diary studies, learners are able to record their impressions or perceptions about learning, unconstrained by predetermined areas of in-

terest. This form of data collection can thus yield insights into the language learning process that may be inaccessible from the researcher's perspective alone. Even in studies in which researchers provide a structure for the diarists to follow (e.g., certain topics to address and guidelines for the content), the researchers are still able to access the phenomena under investigation from a viewpoint other than their own. In addition, because diary entries can be completed according to the participants' own schedules, this approach to data collection allows for greater flexibility than, for example, observations and interviews, which must be scheduled to fit the time constraints of multiple individuals. Bailey's (1983, 1990) influential work offers a classic and complete introduction to the use of diaries in second language research.

6.2.5.2. Caveats

One of the concerns with diary research is that keeping a diary requires a commitment on the part of the participants to frequently and regularly provide detailed accounts of their thoughts about language learning. Because this is often a significant burden to place on study participants, many researchers participate in their own diary studies. However, it is important to note that although the diaries of second language researchers have yielded interesting insights, they constitute a highly specialized population, and the insights from these studies cannot often be extended to other contexts. Another potential complication is that due to the lack of structure of diary entries, data analysis can become a complex affair, making it more difficult for researchers to find and validate patterns in the data.

6.3. ANALYZING QUALITATIVE DATA

In this section, we discuss approaches that are often used to guide the analysis of qualitative data. We also address three important issues in qualitative data analysis—credibility, transferability, and dependability—as well as methods for ensuring that a qualitative study possesses these characteristics.

In analyzing qualitative data, researchers often make use of cyclical data analysis. Basically, this refers to the process of data collection, followed by data analysis, and an hypothesis-formation stage based on the first round of data collection, followed by a second and more focused round of data collection in which hypotheses are tested and further refined, with the process continuing until a rich and full picture of the data

is obtained. Watson-Gegeo (1988, 1997) divided cyclical data analysis into three distinct stages:

- *Comprehensive*: In which all possible aspects of a chosen context are researched.
- *Topic oriented*: In which the topic is clarified through preliminary analysis and focused data collection.
- *Hypothesis oriented*: In which hypotheses are generated based on data.

The hypotheses are then tested through further *focused* and structured interviews, observations, and systematic analysis. In short, cyclical research is the process by which researchers bring increasing focus to their topic of interest.

A similar approach that guides qualitative data analysis is known as grounded theory. This also involves developing theory based on, or grounded in, data that have been systematically gathered and analyzed. Grounded theory attempts to avoid placing preconceived notions on the data, with researchers preferring to let the data guide the analysis. Using grounded theory, researchers often aim to examine data from multiple vantage points to help them arrive at a more complete picture of the phenomena under investigation.

In inductive data analysis the goal is generally for research findings to emerge from the frequent, dominant, or significant themes within the raw data, without imposing restraints as is the case with predetermined coding or analysis schemes. Inductive data analysis is determined by multiple examinations and interpretations of the data in the light of the research objectives, with the categories induced from the data. The framework for analysis is often shaped by the assumptions and experiences of the individual researcher.

6.3.1. Credibility, Transferability, Confirmability, and Dependability

In analyzing qualitative data, researchers must pay attention to three concerns that arise as part of the research: credibility, transferability, and dependability.¹ In terms of credibility, because qualitative research can be

¹In chapter 4, we discussed the importance of internal and external validity to quantitative research. In qualitative research, the notion of internal validity can be related to credibility, and the notion of external validity to transferability. Although they are related, credibility and transferability differ from validity.

based on the assumption of multiple, constructed realities, it may be more important for qualitative researchers to demonstrate that their findings are credible to their research population. Fraenkel and Wallen (2003) suggested several techniques to enhance credibility, including continuing the data collection over a long enough period of time to ensure that the participants have become used to the researcher and are behaving naturally. They also suggested collecting data in as many contexts and situations as possible to make certain that the picture provided in the research is as full and complete as it can be.

For transferability in qualitative research, the research context is seen as integral. Although qualitative research findings are rarely *directly* transferable from one context to another, the extent to which findings may be transferred depends on the similarity of the context. Important for determining similarity of context is the method of reporting known as “thick description,” which refers to the process of using multiple perspectives to explain the insights gleaned from a study, and taking into account the actors’ interpretations of their actions and the speakers’ interpretations of their speech. Davis (1995) distinguished three essential components of thick description:

- *Particular description*: Representative examples from the data.
- *General description*: Information about the patterns in the data.
- *Interpretive commentary*: Explanation of the phenomena researched and interpretation of the meaning of the findings with respect to previous research.

The idea behind thick description is that if researchers report their findings with sufficient detail for readers to understand the characteristics of the research context and participants, the audience will be able to compare the research situation with their own and thus determine which findings may be appropriately transferred to their setting. Other steps can be taken to augment the transferability of research.

For confirmability, researchers are required to make available full details of the data on which they are basing their claims or interpretations. This is similar to the concept of replicability in quantitative research, with the point being that another researcher should be able to examine the data and confirm, modify, or reject the first researcher’s interpretations.

For dependability, researchers aim to fully characterize the research context and the relationships among the participants. To enhance dependability, researchers may ask the participants themselves to review the patterns

in the data. Electronically recorded data can help to recreate the data collection context and allow the researcher to make use of all interpretive cues in order to draw inferences and evaluate the dependability of the inferences that have been drawn. Recordings can also help research participants and other researchers working in similar contexts to assess whether dependable inferences have been derived from the data.

Triangulation involves using multiple research techniques and multiple sources of data in order to explore the issues from all feasible perspectives. Using the technique of triangulation can aid in credibility, transferability, confirmability, and dependability. This important concept is discussed next.

6.3.2. Triangulation

Different types of triangulation have been identified, including theoretical triangulation (using multiple perspectives to analyze the same set of data), investigator triangulation (using multiple observers or interviewers), and methodological triangulation (using different measures or research methods to investigate a particular phenomenon). The most common definition of triangulation, however, is that it entails the use of multiple, independent methods of obtaining data in a single investigation in order to arrive at the same research findings.

As Johnson (1992) noted, “[T]he value of triangulation is that it reduces observer or interviewer bias and enhances the validity and reliability (accuracy) of the information” (p. 146). By collecting data through a variety of means, the researcher helps address many of the concerns with the various qualitative data collection methods that were pointed out earlier in this chapter. One method alone cannot provide adequate support. It may take two or more independent sources to support the study and its conclusions.

For example, in their study of the effects of training on peer revision processes in second language writing, McGroarty and Zhu (1997) assessed the effects of training in terms of students’ ability to critique peer writing, quality of student writing, and students’ attitudes toward peer revision and writing in general. Their experiment included four instructors and 169 students, with each instructor teaching one class in the experimental condition (which included training for peer revision via instructor conferences) and one class in the control condition (which employed peer revision without such training). Their research used a range of different measures, data sources, and methods. The authors pointed out that “the combination of measures, data sources, and methods not only allowed triangulation of the

finding that training for peer revision improves students' ability to critique peer writing and their attitudes toward peer revision, but also illuminated other aspects of peer revision processes" (p. 2).

6.3.3. The Role of Quantification in Qualitative Research

As we saw earlier, some qualitative researchers make use of cyclical data analysis, examining patterns of occurrence in their data and then using them to draw inferences and recursively generate and test hypotheses. Although some qualitative researchers eschew the practice of quantification, others are interested in patterns of occurrence and do not exclude the use of the sorts of numbers and statistics that are usually found in quantitative research. Quantification can play a role in both the generation of hypotheses and the verification of patterns that have been noticed; it can also be used later for the purpose of data reporting.

For example, in Qi and Lapkin's (2001) case study exploring the relationship among quality of noticing, written feedback processing, and revision, they coded two learners' verbal protocols in terms of "language-related episodes" and used quantification to conclude tentatively that higher-proficiency learners may be better able to make use of feedback than lower-proficiency learners. They also suggested that quality of noticing was directly related to L2 writing improvement and that reformulation may be a better technique than error correction in helping learners to notice the gap and produce more accurate language.

Although quantification can assuredly be helpful in the generation of hypotheses and detection of patterns, its practicality is especially evident when the time comes for communicating the findings through publication. As a simple, concise way of reporting general research findings, quantification of some kind is used by many qualitative researchers, who commonly gather enough data to fill a book, and then pare down their data and findings to a length that conforms to journal publication requirements. Quantification is also valuable in that numerical descriptions can make it readily apparent both why researchers have drawn particular inferences and how well their theories reflect the data. Another benefit of quantification is its usefulness to other researchers who may be interested in ascertaining quickly whether the research findings are relevant to other contexts.

6.4. CONCLUSION

In this chapter, we have contrasted qualitative research with other approaches and discussed common methods used for qualitative data collec-

tion. We have also addressed key issues in carrying out qualitative research, including credibility, transferability, confirmability, dependability, triangulation, and quantification. As we have seen, qualitative research can yield valuable and unique insights into the second language learning process. When appropriate qualitative research methods are chosen to address a particular problem, and when the proper standards of empirical rigor are met through triangulation of research perspectives, consideration of emic perspectives, and cyclical data collection and analysis, qualitative research can reliably help us to gain a deeper understanding of the nature of second language learning.

In the next chapter, we turn our attention to classroom research.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Jick (1984) argued that qualitative and quantitative approaches to research should be viewed “as complementary rather than as rival camps” (p. 135). Based on the information you have read in this chapter and your own ideas, why would this comment be valuable to keep in mind in designing a second language case study of two adult learners acquiring English as a second language in their place of employment, an inner-city restaurant kitchen?
2. What is a split (or combination) methods study, and what are some of the advantages of taking this approach?
3. Take a look at the following chart. Based on the information in the chapter and your own ideas, what are the benefits and limitations of each method? Also, in the far right column, provide suggestions for how some of the limitations could be addressed or mitigated.

<i>Method</i>	<i>Benefits</i>	<i>Limitations</i>	<i>Addressing Limitations</i>
Ethnographies			
Case studies			
Interviews			
Observations			
Diaries/journals			

4. Choose three of the methods mentioned in the preceding question. For each, find a research report that makes use of this method. To what extent do you think the method was appropriate to the research question?
5. Provide thumbnail definitions for the terms *credibility*, *transferability*, *confirmability*, and *dependability*. Why are these important for qualitative studies?
6. Imagine that you are a researcher interested in investigating how two children enrolled in a bilingual immersion class at the elementary level respond to corrective feedback provided by their ESL teacher. What method(s) would you employ to gather information about this topic? How would you go about triangulating your findings?
7. What are cyclical data analysis and grounded theory? How do these approaches to qualitative data collection and analysis differ from those taken in quantitative studies?
8. Why might qualitative researchers choose to employ quantification?
9. In both qualitative and quantitative studies, it is important to discuss the role of the researcher in the data collection process. However, why would this be especially important in a qualitative study?
10. Think about your own research interests for a moment. If you were considering taking a qualitative approach to investigating your topic, how would you begin your study? In particular, think about your research question and how it might change over the course of gathering data, what methods you would use to collect data, and what conclusions you could draw from those data.

Classroom Research

This chapter addresses common practices in classroom-based research. We begin with a description of classroom observation techniques followed by discussion of a range of observation schemes. We then present three commonly used introspective measures followed by a discussion of some of the practical and logistical considerations involved in carrying out research in L2 classrooms.¹ Finally, we move on to a description and discussion of methodology in two areas of classroom research: the role of instruction in second language development, and action research.

Although classrooms constitute a distinct *context* for research, many of the methodological practices and data collection techniques associated with classroom research are not unique to classroom settings, and some are also discussed elsewhere in this book. For example, we discuss diary studies both in chapter 6 as part of qualitative research methods, and in the current chapter where we focus exclusively on diary use by learners and teachers in second and foreign language classroom contexts. We begin the chapter with a discussion of the nature of classroom research.

7.1. CLASSROOM RESEARCH CONTEXTS

Traditionally, second language researchers have distinguished between classroom-based research and research conducted in controlled laboratory contexts. Typical laboratory-based research has the advantage of allowing the researcher to tightly control the experimental variables, randomly assign subjects to treatment groups, and employ control groups—all of

¹We use the term *second* in conjunction with *language* to convey third, fourth, bilingual, foreign, and so on throughout this book. Likewise, we often use the term *second language* in conjunction with *classrooms* to include foreign language classrooms.

which are difficult, and sometimes impossible, to implement in classroom-based research contexts. Such concerns regarding classroom research have led some second language researchers to claim that although laboratory settings are more abstract, the benefits connected with being better able to control and manipulate intervening variables may be worth the potential costs of abstraction (Hulstijn, 1997).

Whether research carried out in the laboratory can (or cannot) be generalized to the L2 classroom is an empirical question. In any case, in light of the complementary strengths and limitations of laboratory and classroom studies, second language researchers are increasingly recognizing that studies must be carried out in different contexts and that a range of different approaches must be used to gain a deeper understanding of the complexity of second language learning. Thus, whereas classroom research can enhance our understanding of how to implement effective ways of improving learners' second language skills, laboratory studies can provide more tightly controlled environments in which to test specific theories about second language development.

Combined approaches to classroom research—that is, those involving a range of different approaches, including both experimental and observational techniques—are also gaining popularity. Allwright and Bailey (1991), for example, pointed out that “increasingly it appears, second language classroom researchers are calling for judicious selection and combined approaches rather than rigid adherence to one approach over another” (p. 68). We likewise suggest that research in a wide range of contexts and using multiple methods and techniques will be necessary for developments in the ongoing investigation of how second languages are learned and, consequently, how languages may best be taught. We now turn to a discussion of some of the common data collection techniques used in research set in second language classrooms.

7.2. COMMON TECHNIQUES FOR DATA COLLECTION IN CLASSROOM RESEARCH

7.2.1. Observations

As we discussed in chapter 6, observational data are common in second language research and observations are a useful means for gathering in-depth information about such phenomena as the types of language, activities, interactions, instruction, and events that occur in second and foreign lan-

guage classrooms. Additionally, observations can allow the study of a behavior at close range with many important contextual variables present. Here, we focus on the particular concerns that can arise when carrying out observations in intact classrooms, as well as providing information about the different types of observation schemes that have been developed by second language classroom researchers.

7.2.1.1. Conducting Classroom Observations

Obtrusive Observers. Observation etiquette may initially seem secondary to the more practical nuts and bolts of carrying out a thorough observation, but conforming to good observation etiquette is very important. Any observer in the classroom runs the risk of being an obtrusive observer, which can be problematic for research. An obtrusive observer's presence may be felt in the classroom to the extent that the events observed cannot be said to be fully representative of the class in its typical behavior, and therefore the observation data may have limited validity. (We also discussed this in chap. 6 as the observer's paradox).

An obtrusive observer may also be problematic for the instructor and students in terms of compromising the quality of the lesson, preventing instructors from delivering the lesson to the best of their ability and, consequently, preventing the students from learning to the best of theirs. For example, younger learners in particular can become very easily distracted by observers. They may be interested in the recording equipment and may pay more attention to a new person in their classroom with a digital video or voice recorder than to their instructor.

The Hawthorne Effect. Another potential problem for observational research is the so-called Hawthorne effect, also discussed in chapters 4 and 6. This effect was first described by observers at the Hawthorne, Chicago branch of the Western Electric Company (Brown, 1954; Mayo, 1933). When the observers were present, the productivity of workers increased regardless of whether or not there were positive changes in working conditions. The workers were apparently happy to receive attention from researchers who expressed an interest in them by observing them, and this impacted their behavior. Accordingly, whereas in observational research it may be difficult to be sure that the observed classes are the same as they would be without the observation, in controlled research it may be difficult to separate Hawthorne effects from experimental variables. Although Hawthorne effects in management have been queried by some (e.g., Adair,

1984), including those who have pointed to the small number of participants in the original study, educationalists usually make efforts to take such effects into account when conducting observations. For example, as mentioned in chapter 5, Mellow et al. (1996) argued that time-series designs are particularly useful for investigating and evaluating different approaches to second language teaching, pointing out that one of their benefits is that they may reduce the Hawthorne effect as students and teachers begin to feel more comfortable and natural about being observed.

Objectivity and Subjectivity. Classroom observations are not only conducted by researchers external to the school or educational environment. Instructors often observe each other's classes for professional development as well as for research purposes, and they may also carry out observations of their own classes, usually using audio or videotapes to assist with this process. This brings us to another consideration that needs to be taken into account when conducting observations: namely, the level of objectivity or subjectivity of the observer. The level and impact of this on the study is often debated and needs to be clearly recognized and reported in research. Whereas objectivity is typically valued in second language research, particularly in experimental work, both objectivity and subjectivity have their respective roles in research on second language learning. Therefore, in classroom studies, it is necessary for researchers to both strive for objectivity and also be aware of the subjective elements in that effort—for example, in how they gather data, analyze data, and report the results of analyses.

Obtaining Permission to Observe and Enlisting the Help of the Instructor. In addition to keeping issues of objectivity and subjectivity in mind, there are several further precautions for observers to take into account. First of all, it is important to obtain the permission of the instructor to observe the class well in advance of the scheduled observation(s). This is not only a professional courtesy, but may also help the instructor to lessen any impact of the observation on lesson planning and implementation. When working with schools and language programs, the researcher should not assume that the permission of administrators indicates that individual classroom instructors have been informed and that their instructional schedules have been considered. It is important to contact the instructor in advance in order to obtain consent and to negotiate the schedule and observation process. It is also important to seek the instructor's input about matters such as when to arrive. For example, arriving a little before the learners,

or at the same time, or during an activity when their attention will be focused elsewhere, can all be options for lessening the impact of the observer. The instructor may also have the best idea about where the observer can sit in the classroom so as to be minimally intrusive. Some instructors prefer observers to sit in the back or off to the side of the class, and some may recommend that researchers begin coming to class several days before they conduct the research in order to habituate the students to their presence. Instructors may feel that by minimizing the presence of the observer, their students will not become distracted by note taking or the direction of the observer's gaze or equipment, and will concentrate on the lesson.

For some classes, students may be used to the occasional presence of a supervisor or instructor-trainer, and little explanation will need to be provided. However, in other classes, especially those early in the semester or program, students may never have experienced an observer before. For these classes, depending on the research problem being studied, it might be possible to introduce the researcher into the classroom so that they participate in the instructional activities and are seen as an instructor's aide. Another reason to be sure to have made personal contact with the instructor beforehand is that having more than one observer in the classroom at any time could be disruptive for the instruction, and the instructor is usually in the best position to know the schedule for who plans to observe when. If the observation is to be ongoing, it may also be wise to ask the instructor for feedback after the class in case they would prefer something to be done differently the next time around. Murphy (1992) recommended that observers who are there for research purposes keep in mind that their role is not to judge, evaluate, criticize, or offer constructive advice, and also that if asked by the learners what they are doing in the classroom, observers keep their responses as short as possible.

Debriefing the Instructor. It is also important as part of the negotiation surrounding the observation process to debrief the instructor about the research findings or the content of the observation notes or scheme. Timing is also an important consideration here. For example, researchers might provide instructors with a copy of their notes after each lesson or arrange a time to meet in order to discuss the research. By keeping the observation process as transparent and interactive as possible, researchers can often establish a more trusting and cooperative relationship with instructors. Of course, in some cases, the instructors may be the focus of the research, or it may unduly influence the research if they are kept continually

debriefed. In these cases, it may be preferable to make such contact after the project has been completed.

Finally, of course, it is always important to thank the instructor for allowing the observation, together with any other parties who have been helpful to the research, including both administrators and students. It can be easy to overlook such simple things, but in fostering good relationships between instructors and future researchers, expressing courtesy and appreciation is important. As part of this process, when research is published with acknowledgments made to schools, principals, instructors, and students, it is helpful to send copies of the publication to the schools and instructors because they may not have access to the same journals and publications as the researcher. Following is a helpful checklist to consider in setting up observations:

- Contact the classroom instructor (in person if possible).
- Determine the schedule for observation.
- Negotiate the observer's role in the classroom, including regular previsits, arrival time, introductions, and seating arrangements.
- Debrief the instructor (either during or after the observational period) on the findings of the study.
- Clearly express appreciation to the instructor, students, and administration.

7.2.1.2. Observation Procedures and Coding Schemes

When considering observation procedures and coding schemes, the first critical step is to carefully consider the goals of the research and the observation. If an existing observation procedure or coding scheme can be used or adapted, this can prevent duplication of effort in developing new schemes. Readily available observation schemes address a wide range of classroom phenomena, and a number of observation schemes have been developed for researchers working in L2 classrooms (schemes have been developed by Allen, Fröhlich, & Spada, 1984; Fanselow, 1977; Mitchell, Parkinson, & Johnstone, 1981; Moskowitz, 1967, 1970; Nunan, 1989; Sinclair & Coulthard, 1975; Ullman & Geva, 1983). Existing coding schemes can vary considerably in their organization and complexity, ranging from simple checklists and tallies of behaviors to highly complex schemes for judging the meaning or function of particular behaviors, as well as combination schemes. In their book focusing on the second language classroom, Allwright and Bailey (1991) provided examples of seven

different published coding schemes. Other useful discussions and examples have been supplied by Chaudron (1988), Lynch (1996), and McDonough and McDonough (1997).

Description of Observation Schemes. In most observation schemes, the observer marks the frequency of an observed behavior or event at a regular time interval; for example, observations may be made of every instructor question or of the students' reactions to the writing on the board every 5 minutes. In Nunan's (1989) classroom observation tally sheet, for instance, there are categories for such classroom events as the instructor's praise, instructions, and explanations of grammar points, as well as learners' questions, answers, and interactions with other students. Categories such as Nunan's are considered low inference; that is, "clearly enough stated in terms of behavioral characteristics . . . that observers in a real-time coding situation would reach high levels of agreement or reliability" (Chaudron, 1988, pp. 19–20). Nunan's scheme appears in Table 7.1.

Other observation schemes incorporate both low-inference and high-inference categories. High-inference categories are those that require judgments, such as in relation to the function or meaning of an observed event. For example, the Target Language Observation Scheme (TALOS; Ullman & Geva, 1985) consists of two parts. The first is a real-time, low-inference checklist for describing live classroom activities (e.g., drills, dialogues, translation, free communication), linguistic content (e.g., sounds, words, phrases, discourse), and skill focus (e.g., reading, writing, listening, speaking), as well as teaching behaviors (e.g., drills, narrations, explanations, comparisons, answers, discipline) and student actions (e.g., types of questions asked). The second part of the observation scheme is a high-inference rating scale to be completed after the observation. Here, the observer provides ratings on a 5-point scale (extremely low to extremely high) for categories such as enthusiasm, humor, and negative and positive reinforcement. The TALOS scheme appears in Fig. 7.1.

The Communicative Orientation of Language Teaching, or COLT, is a similar sort of structured observation scheme (Allen et al., 1984). Developed in the 1980s to describe differences in communicative language teaching, the COLT scheme focuses on pedagogic and verbal behavior in two sections—one section for real-time coding (Part A) and the other section (Part B) for postobservation analysis of tape recordings. In Part A, more than 40 categories are provided for participant organization and activities, as well as topic type, content, and control. In Part B, the observer is provided with a chart that allows for post-hoc analyses of student-instructor

TABLE 7.1
Classroom Observation Tally Sheet from Nunan (1989).

	<i>Tallies</i>	<i>Total</i>
1. Teacher asks a display question (i.e. a question to which she knows the answer)	/ / /	3
2. Teacher asks a referential question (i.e. a question to which she does not know the answer)	/ / / /	4
3. Teacher explains a grammatical point		0
4. Teacher explains meaning of a vocabulary item		0
5. Teacher explains functional point		0
6. Teacher explains point relating to the content (theme/topic) of the lesson	/	1
7. Teacher gives instructions/directions	/ / / / /	6
8. Teacher praises	/	1
9. Teacher criticises		0
10. Learner asks a question	/ / /	3
11. Learner answers question	/ / / /	4
12. Learner talks to another learner		0
13. Period of silence or confusion		0

Source: Nunan, David, *Understanding language classrooms: A guide for instructor initiated action*, 1st Edition, © 1989. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.

and student–student interaction within various activity types. This observation scheme has been used in original or modified forms in a wide range of classroom studies (e.g., Lightbown & Spada, 1990, 1994; Lyster & Ranta, 1997; Spada & Lightbown, 1993; White, Spada, Lightbown, & Ranta, 1991). Figure 7.2 shows the COLT scheme.

In Lyster and Ranta's (1997) study of corrective feedback and learner uptake, they examined four immersion classrooms, with their transcripts totalling 18.3 hours of classroom interaction. They noted that:

although the teachers knew we were interested in recording classroom interaction, they were unaware of our research focus related to classroom

feedback. The teachers continued with their regular program while recordings were being made, and one or more observers coded classroom activities using Part A of the Communicative Orientation to Language Teaching coding scheme (Spada & Fröhlich, 1995), which we had adapted for use in immersion classrooms. Because we were interested in analyzing teacher behaviour in this first phase of a larger program of research, we focused exclusively in our analyses on teacher–student interaction (p. 43)

This extract illustrated Lyster and Ranta's (1997) research process. First, they identified a helpful starting point for their observation scheme, the COLT. Then they adapted it to their particular immersion classroom context. Finally, they narrowed their focus to teacher–student interactions only.

Although classroom observation coding schemes vary considerably, some common elements may be identified. For example, many schemes include a category relating to the identity of the participants and their groupings (e.g., small or large groups). Most schemes also have categories for the

General Information

1. Observer	_____			
2. Date	_____			
3. School	_____			
4. Name of French teacher	_____			
5. Name of home-room teacher	_____			
6. Grade	_____			
7. Observation (circle one)	1st	2nd	3rd	4th
8. Lesson start	_____			
9. Lesson end	_____			
10. French room (circle one)			yes	no
11. Teacher French Displays?			yes	no
12. Student French Displays?			yes	no
13. Francophone French teacher?			yes	no

FIG. 7.1. The TALOS observation scheme. Source: Ullmann, R., & Geva, E. (1985). Expanding our evaluation perspective: What can classroom observation tell us about Core French Programs? *The Canadian Modern Language Review*, 42(2), pp. 319–322. Reprinted by permission of *The Canadian Modern Language Review Inc.* (www.utpjournals.com).

(continued on next page)

The High Inference TALOS					
<i>Teacher</i>	extremely low	low	fair	high	extremely high
Use of L ₁	0	1	2	3	4
Use of L ₂	0	1	2	3	4
teacher talk time	0	1	2	3	4
explicit lesson structure	0	1	2	3	4
task orientation	0	1	2	3	4
clarity	0	1	2	3	4
initiate problem solving	0	1	2	3	4
personalized questions & comments	0	1	2	3	4
positive reinforcement	0	1	2	3	4
negative reinforcement	0	1	2	3	4
corrections	0	1	2	3	4
pacing	0	1	2	3	4
use of audio-visual aids	0	1	2	3	4
gestures	0	1	2	3	4
humour	0	1	2	3	4
enthusiasm	0	1	2	3	4
	0	1	2	3	4
<i>Students</i>					
Use of L ₁ on task	0	1	2	3	4
Use of L ₂ on task	0	1	2	3	4
student talk time on task	0	1	2	3	4
initiate problem solving	0	1	2	3	4
comprehension	0	1	2	3	4
attention	0	1	2	3	4
participation	0	1	2	3	4
personalized questions & comments	0	1	2	3	4
positive effect	0	1	2	3	4
negative effect	0	1	2	3	4
S to S interaction on task	0	1	2	3	4
	0	1	2	3	4
<i>Program</i>					
linguistic appropriateness	0	1	2	3	4
content appropriateness	0	1	2	3	4
depth	0	1	2	3	4
variety	0	1	2	3	4
listening skill focus	0	1	2	3	4
speaking skill focus	0	1	2	3	4
reading skill focus	0	1	2	3	4
writing skill focus	0	1	2	3	4
formal properties	0	1	2	3	4
functional properties	0	1	2	3	4
integration with general curriculum	0	1	2	3	4

FIG. 7.1. (continued)

content or topic of the lesson, as well as the types of activities and materials used. There may also be categories dealing with the language employed during an activity or event (e.g., L1 or L2) and the targeted skill. Depending on the researcher's particular goals and questions, the scheme may additionally include categories for marking the frequency or duration of targeted activities or behaviors. Observation schemes may be used to code broad categories related to classroom instruction, or they may be focused on specific characteristics of a single classroom phenomenon.

As illustrated in Fig. 7.3, there are many possible categories that could be included in even the most focused observation scheme, and it is unlikely that any one scheme could capture all the potentially relevant aspects of information about classroom events. Figure 7.3 shows a sample observation scheme that could be used in a study of feedback in second language classrooms. There is increased interest, mostly in ESL classes, in the presence and benefits of feedback in classroom settings (e.g., R. Ellis, 2000; Mackey, 2000; Oliver, 2000; Samuda, 2001; Williams, 1999). One way to code data is to use a scheme, such as that depicted in Fig. 7.3, to determine the frequency of the provision and use of feedback by instructors and students in language classes. The highly structured nature of the scheme would also allow the researcher to compare the frequency of feedback types, feedback focus, and uptake in classes at different institutions or among learners with different ages or different language abilities.

By making tallies and notes on the sample observation sheet given in Fig. 7.3, a researcher could record the sources and types of feedback, the linguistic objects of the feedback (i.e., the errors), and what sort of uptake has occurred, if any. There is space for examples of any of these categories, together with space to note how many times they occur in units of 10 minutes.

The following list briefly notes some of the advantages of using or modifying existing observation schemes. (A more in-depth discussion of some important caveats follows in the next section.)

- Relative ease of use when compared with nonsystematic classroom descriptions with no preexisting guidelines or descriptions of data based on the schemes.
- Comparability with other studies, with a potentially concomitant increase in the generalizability of the research.
- Simplified analysis of complicated and rich, but possibly overwhelming, classroom data.
- Possibility of measuring change or status over different time periods.

vant to the research questions. For example, a scheme focused on instructor–learner dialogues may not allow a researcher to adequately characterize the nature of language use in the classroom; without adaptation, it would not be appropriate if the research question focused primarily on small-group dynamics in the L2 classroom. Observation schemes can promote valid findings only when they are appropriate and applicable to the research question. Researchers should also consider the type of findings that are likely to emerge from an observation scheme. If the research question goes beyond descriptions of behaviors, for instance, an observation scheme based on low-inference categories is not likely to highlight the items of interest in the data.

Another consideration is the use of time as a unit in an observation scheme. If the overall occurrence of a phenomenon is of interest (e.g., how many times an instructor recasts learner utterances), then a category system like the COLT is most appropriate. In a category system, the observer checks a behavior each time it occurs in order to record its frequency. Alternatively, if the distribution of a phenomenon throughout the class is of interest, the observer can employ a sign system, in which an observation is made at regular intervals of time. Also, it is important to note that unless more than one observer is present in the classroom, or the data are videotaped and later replayed for a second person, with most coding schemes only one rater observes (and at the same time codes) the data. We discuss issues of coding and interrater reliability in chapter 8.

By helping to ensure not only that relevant aspects of the classroom lesson are noted and remembered, but also that significant patterns of interaction are identified, observation schemes have allowed researchers to gain a deeper understanding of the inner workings of second language classrooms. However, some researchers have argued that the use of observation schemes with their predetermined categories “seriously limits and restricts the observer’s perceptions—that it creates a kind of tunnel vision because the observer sees only those behaviors that coincide with the categories in the observation scheme” and may fail to observe other important features (Spada, 1994, p. 687). Essentially, as we discuss in chapter 8, the coding scheme may result in data reduction, where potentially important patterns are missed. As one simple example, making only observations at 5 minute intervals ignores the potentially rich events that occur in between. In addition, even the most thorough observation schemes cannot allow the researcher to reach conclusions about what the participants themselves are experiencing. Another criticism leveled at observation schemes is that there

is insufficient evidence showing that the categories are valid predictors of learning processes and outcomes. One way to address such criticisms may be to record the classroom data and then to develop custom-made coding schemes based on the observations, with the coding developing out of the observations. (Custom-made schemes are further discussed in chap. 8, and coding in which categories emerge from the data were discussed in chap. 6.) Whether customized or preexisting schemes are used, additional data-gathering methods may be helpful in order to triangulate classroom data and provide multiple perspectives by accessing the learners' insights into the events that have been observed.

7.3. INTROSPECTIVE METHODS IN CLASSROOM RESEARCH

Introspective methods—or data-elicitation techniques that encourage learners to communicate their internal processing and perspectives about language learning experiences—can afford researchers access to information unavailable from observational approaches. In second language research, a range of introspective methods have been employed. These methods vary with respect to the practicality of their application to classroom research. Uptake sheets, for example, described in the next section, allow researchers to investigate learners' perceptions about what they are learning. Stimulated recalls (see also chap. 3) may yield insights into a learner's thought processes during learning experiences, whereas diaries (discussed in chap. 6) can present a more comprehensive view of the learning context from a participant's viewpoint. The following discussion provides an overview of the use of some introspective methods that are particularly relevant in second language classrooms and addresses the advantages and applicability of each.

7.3.1. Uptake Sheets

One way to elicit learners' perspectives on second language classroom events is through the use of uptake sheets. Uptake sheets were initially developed as a method of data collection following Allwright's (1984a, 1984b, 1987) interest in learners' perceptions about what they learned in their language classes. He collected learners' reports about their learning, which he termed *uptake* or "whatever it is that learners get from all the language learning opportunities language lessons make available to them" (1987, p. 97). In classroom research, uptake sheets are often distributed at the beginning of the lesson, and

learners are asked to mark or note things on which the researcher or teacher is focusing. Whether used to uncover information about learning, noticing, attitudes, or a range of other interesting phenomena, uptake sheets can allow researchers to compare their own observations and other triangulated data with information obtained from the learners, and they create a more detailed picture of classroom events in the process.

An example of an uptake chart can be found in Fig. 7.4. This sheet is designed for classroom learners to fill out during a lesson or activity by a teacher or researcher who wanted to elicit information about what learners were noticing about second language form.

In their study of the effects of different uptake sheet formats on learner reports about their learning, Mackey, McDonough, Fujii, and Tatsumi (2001) asked learners to mark uptake sheets in order to address research questions focusing on the relationship between the format of the uptake sheet and the quantity and quality of learner reporting. Learners were asked on all three formats to indicate “(a) which language forms or concepts they noticed, for example, pronunciation, grammar, vocabulary, or

What are you noticing about...	Who said it? (check as many as you want)				Was this new to you?		
	Teacher	Classmate	Me	Book	Yes, new	No, heard of it.	No, knew it.
<i>Pronunciation</i> • • • • • •							
<i>Vocabulary</i> • • • • • •							
<i>Grammar</i> • • • • • •							

FIG. 7.4. Sample uptake chart.

business; (b) who produced the reported items, for example, the learner, the instructor, or their classmates; and (c) whether the reported items were new to the learner” (Mackey et al., 2001, p. 292). The learners in this study were given sheets at the beginning of class for each of six consecutive classes and asked to fill them out as they noticed language forms or concepts during the instruction.

7.3.2. Stimulated Recall

Another way to obtain the perspectives of classroom learners is through stimulated recall. In this method used in the classroom, described more generally in chapter 3, the observer makes an audiotape or videotape of a lesson for the stimulus, and then plays the tape to a participant, periodically stopping the tape to ask what the participant had been thinking at that particular point in time. Stimulated recall can be used to provide the researcher with access to the learners’ interpretations of the events that were observed and can be a valuable source of information for researchers interested in viewing a finely detailed picture of the classroom. A detailed account of stimulated recall methodology, as well as considerations in applying it in classroom and laboratory studies, can be found in Gass and Mackey (2000).

Stimulated recall has been used to investigate various aspects of second language classrooms, as exemplified in Roberts (1995). Roberts used stimulated recall in a study of learners’ recognition of feedback in a university Japanese as a Foreign Language class. He recorded a 50-minute class period, which was viewed several days later by three volunteers from the class. The participants were asked to write down their perceptions about episodes from the tape involving instructor feedback and to note the error being corrected. This study exemplifies one of the contributions that stimulated recall can make to classroom research by allowing researchers to view instruction from the learners’ perspectives.

7.3.3. Diary Research in Classroom Contexts

Bailey (1990) defined a diary study as “a first person account of a language learning or teaching experience, documented through regular candid entries in a personal journal and then analyzed for recurrent patterns and salient events” (p. 215). Diaries of classroom contexts can produce useful data on a range of aspects of the second language learning process. These include individual learners’ and instructors’ insights into their own learning

and teaching processes, their self- and other-comparisons, decision-making processes, the process of development (or not) over time, attitudes toward classroom learning and teaching, the use of strategies, and the recognition and use of feedback. Diary studies have the additional advantage of time sensitivity. Because most diary research is longitudinal, it can illuminate how perceptions develop over time. However, as discussed in chapter 6 on qualitative research, there are also some drawbacks to the use of diaries, including the highly subjective nature of the data.

In classroom research, some structure can be provided for the diary entries. It is increasingly the case that instructors ask learners to keep diaries as a part of coursework (and even course assessment), for which the goal is enhancement of pedagogy. The diaries might be required to address specific points, including how well the learners have followed specific lessons; what is enjoyable, easy, or difficult about the instruction; and their reactions to the instructor and other learners, as well as to specific classroom activities and group and pair work.

Instructor diaries are also common in educational research, and explorations of language instructor diaries are becoming more common in the second language research field. For example, Bailey et al. (1996) used instructor diaries to investigate the role of language-learning and -teaching beliefs in decision making by student instructors. An examination and comparison of the instructors' diaries indicated that conscious examination of long-held beliefs about language learning helped to shape pedagogical decisions, and that student teaching was experienced differently according to the student instructor's gender, educational background, and language-learning experiences. In general, instructors' diaries have tended to focus on classroom experiences, perceptions about student reactions and learning, and instructional decision making (and decision changing) for which the method matched the goals of the research.

Diary research represents a significant expenditure of time, both for those who write the diaries as well as for the researchers who analyze them. When embarking on diary research, it is important for diary writers to schedule regular times for writing. The quality of the diaries can also be enhanced if the researcher includes guidelines for the range and amount of writing expected per entry or provides sample questions that the writer may want to consider for each entry. Diary writers should be encouraged to keep a notebook, mini audio recorder, or personal digital assistant with them to jot down insights as they occur and transfer them later to the diary (or leave them on tape / disk in the case of oral diaries). They should also be

reminded to include examples to illustrate their insights in the diary entries and can be asked to include their own questions about teaching and learning that occur as the research progresses.

Bailey's (1983) survey of journal studies and discussion of her own journal describing her learning of French illustrated such phenomena as the role of self-esteem in second language learning, operationalized as competitiveness and anxiety, as in the following quotations:

I feel very anxious about this class. I know I am (or can be) a good language learner, but I hate being lost in class. I feel like I'm behind the others and slowing down the pace (pp. 75–76)

Today I was panicked in the oral exercise where we had to fill in the blanks with either the past definite or the imperfect. Now I know what ESL students go through with the present perfect and the simple past. How frustrating it is to be looking for adverbial clues in the sentence when I don't even know what the words and phrases mean. I realized that the teacher was going around the room taking the sentences in order so I tried to stay one jump ahead of her by working ahead and using her feedback to the class to obtain confirmation or denial of my hypotheses. Today I felt a little scared. I'm so rusty! (p. 74)

The analysis of diary data involves a careful and thorough search for patterns in the writing or tapes in order to find recurrent themes of interest. When reviewing the data, it is important to be conscious of one's own beliefs, experiences, and orientations to the question of interest and how these may influence interpretations both of individual diary entries and of the emergent patterns. When diaries from several participants are included in the research, it is important to note not only how many times a phenomenon is noted, but also by whom it is noted. This will help researchers to avoid seeing the experiences of only a few participants as being reflective of the experiences of the whole. Similarly, the salience of the phenomena in the diary entries should be considered in order to prevent decontextualized over- or underemphasis of the previously mentioned points.

7.4. PRACTICAL CONSIDERATIONS IN CLASSROOM RESEARCH

Studies carried out in second language classrooms take a wide variety of forms, ranging from ethnographic work on 'naturalistic' classroom discourse or interaction to quasi-experimental studies of the effects of specific

instructional practices such as form-focused instruction, extensive reading, and processing instruction. Classroom research has enhanced our understanding of second language learning in a variety of contexts, including second and foreign language contexts, as well as in classes with differing orientations to language teaching. It contributes in important ways to our understanding of both second language learning and second language teaching. However, the process of carrying out classroom research is both complex and time consuming. The purpose of this section is to detail some of the considerations, logistical and conceptual, that researchers should weigh when designing classroom studies.

7.4.1. Logistical Issues to Consider When Carrying Out Classroom Research

Together with decisions about particular data collection methods to use in obtaining data on second language classroom contexts, researchers also need to consider matters of logistics. As mentioned earlier, conducting observations in classrooms raises a unique set of concerns and issues, and indeed logistical matters in classrooms are quite different from those in laboratories (see papers in Schachter & Gass, 1996, for further discussion).

When observing a classroom, it is common not only to use field or observation notes and/or a coding scheme, but also to triangulate or supplement this method with a mechanical means of recording the lesson, such as audio or video recording. In a laboratory research setting, recording can be a relatively straightforward matter, although the quality required of the recording will differ, of course, depending on the emphasis of the research. However, in classrooms, recording can present a unique set of problems.

First of all, in selecting the most appropriate recording device, the particular nature of the data collection should be considered. For both audio and video recordings, digital and analog (tape) recorders are now available. Digital recordings tend to have higher-quality sound, and digital files can be manipulated for analysis and presentation more easily. Digital recordings also tend not to degrade over time. However, most transcription machines are made for cassette tapes, and depending on whether or not software is available, digitally recorded data can be more expensive to transcribe.

Whether using digital or tape recordings, the use of microphones also needs to be considered, as mentioned in chapter 3. If a participant can speak directly into the recorder, the internal microphone may be sufficient. This may be less distracting and more foolproof, because less equipment (and

therefore less possibility for equipment malfunction) is involved. However, if the goal is to capture small-group or pair work, or a number of individuals in a classroom, external microphones are often needed. One type of external microphone, the lapel microphone, can make it easier to distinguish the wearer's voice from other voices on a recording. Boom microphones, on the other hand, can capture the speech of several participants. The selection of a microphone will also depend on the nature of the research. If the data collection calls for recording a large-whole class activity, the most sensitive microphone possible should be used. However, if the research involves recording separate, simultaneous group activities, using a very sensitive microphone might pick up talk from both the target group and adjacent groups, making it more difficult to transcribe and analyze the group discussion. In general, if the research question allows grouping or pairing of learners into male/female sets, this can make transcription easier.

The research questions may also suggest the appropriateness of videotaping (e.g., a study on nonverbal communication). Even if only audio-taping is required, it can be very useful to have a supplementary videotape—to check nonverbal signals, for example, or as a backup for the audio, to aid in transcription, and so on. When using a video camera, the focus of the study needs to be carefully considered. More than one camera might be necessary to capture student interactions in small-group and pair work as well as instructor input. However, more than one camera can require more than one operator, and this can double the intrusion into the classroom. Sometimes classrooms are equipped with concealed cameras that allow remote operation. This is the exception rather than the rule, however, and it is important to note that fully informed consent must always be sought. If only one camera is available, it can help to try to place it in a corner of the room so that not only the instructor but also as many of the students as possible are captured on tape. In this way, more information can be gathered on the interactions between the instructor and the students and among the students themselves.

When using audio recorders, researchers should keep in mind that the quality of these can vary extensively. As a general rule, it is important to use more than one as a backup in classroom research. This is both to pick up more data and to account for equipment failure or human intervention such as learners turning equipment on and off and so on. It might also be possible to position the recorder near the instructor and place microphones (if available) at various locations in the classroom to pick up the students' voices. Movable chairs and tables are useful in this regard; if there are only a

limited number of recorders or microphones, the researcher can arrange the students around whatever equipment is available.

When recording younger children, as noted earlier, researchers must also keep in mind that the equipment will be novel, interesting, and thus a target for fiddling. One way to help ensure that the equipment will remain intact is to ask for colleagues or adult volunteers to come to the class and keep an eye on it. However, this may be disruptive for the class and alter its nature. Alternatively, a researcher could start bringing the equipment to the class a few weeks before the observation. In that way, the children could become accustomed to the presence of both the researcher and the equipment.

In any class, there will also most likely be learners who have not consented (or whose parents have not consented for them) to be videotaped. In such cases, it is necessary to make arrangements for these learners to sit behind the video camera so that they are not recorded, and to be kept away from any microphones that may be placed in the class.

Other logistical concerns relate to the physical environment of the classroom. In some data-gathering situations, it will be important to ascertain whether or not the chairs and tables are movable, to determine the quality of the acoustics in the classroom, and to note the availability of writing implements and boards. Issues such as the temperature, light, and the scheduling of breaks can also impact data collection. The following checklist is helpful to consider when working out the logistics of classroom research:

- Select a recording format that will facilitate the ultimate uses of the data (e.g., transcription, analysis, presentation).
- Consider whose voices and actions need to be recorded, as well as how sensitively and distinguishably this needs to be done and in which situations.
- Determine what kinds of microphones and other equipment should be used for these purposes and where they should be placed to collect as much relevant data as possible.
- Supplement your primary recording method with a backup, but try to gauge what is necessary and sufficient for the job in order to avoid equipment malfunction or undue complexity. Pilot testing can help.
- Consider the amount of intrusion in the classroom caused by equipment and equipment operators.
- Take anonymity concerns seriously and act accordingly.
- Plan the physical arrangement beforehand, taking into account the suitability and adaptability of the environment.

- Consider human factors such as the age of the participants and how the equipment may affect them; acclimate participants if necessary.

7.4.2. Problematics

In addition to the logistical concerns arising out of classroom observations, there are other issues that need to be addressed when conducting classroom-based research. In this next section, we discuss several of the most serious, including dealing with relevant nonparticipating parties (i.e., parents and administrators), debriefing participants, accessing test scores if necessary, and segmenting the data for presentation.

7.4.2.1. Informed Consent

As detailed in chapter 2 in the section on informed consent, it is necessary to obtain the consent of all interested parties when conducting research. In the case of classroom-based research, this usually means that consent must be obtained from learners (and their parents if the learners are children), the instructor, and the school administrators. As in all research, all parties must be informed as to the purpose of the research and what participation entails. In classroom research it is particularly important that potential participants do not feel pressured by their instructors and are assured that non-participation in the research will incur no penalties. They must also be assured that every effort will be made to accommodate those individuals who do not wish to participate by not using any data inadvertently collected, such as their voices on tape.

7.4.2.2. Debriefing Participants and Facilitators

It is also important that participants—especially parents, instructors, and administrators—be debriefed after the conclusion of the study. Researchers may, for example, wish to send a letter to the parents detailing the results of the study. For administrators, a researcher may also consider setting up a meeting to discuss the results of the research. In this way, administrators can be assured that the research was beneficial and a worthwhile use of their limited time and resources. Although research may not have direct applications that benefit a particular school or language program, any well-designed and well-motivated research project should enhance our understanding of the nature of second language learning, and

it may therefore have eventual benefits for second language teaching. Many instructors and administrators appreciate knowing how individual research projects conducted in their school context fit into the theory and practice of second language research.

7.4.2.3. Ensuring Confidentiality and Minimizing Disruption

In addition to participant issues, the classroom researcher will also face concerns related to the gathering of data. To triangulate the findings from classroom observations, for instance, the researcher may also wish to obtain test scores or report cards. It is important to remember that student grades are a highly sensitive matter. Permission to view grades or any graded material should be included in the informed consent and discussed with both instructors and administrators, and participants should be assured of confidentiality. If researchers prefer to administer their own tests, arrangements will need to be made so that the testing does not take away more class time than absolutely necessary. Indeed, it is always important to consider the most judicious use of class time when conducting classroom research. Researchers need to be sensitive to the perspectives of both instructors and learners in the classroom, and they should be careful not to disrupt learning during the research whenever possible.

7.4.2.4. Data Segmentation and Coding

As we discuss in chapter 8 on coding, once the data have been gathered, the researcher must also decide how to segment the data for presentation. It is the responsibility of the researcher to analyze and present the data in a manner that will be accessible to interested parties. When determining units of analysis for classroom data, it is important to consider both the aims of the research and the classroom context. With regard to research questions, data should be analyzed and presented in ways that can shed light on the specific questions asked. For example, if the researcher is interested in the quantity of talk by learners in groups or pairs, one appropriate unit of analysis might be the word. If, on the other hand, the researcher wants to investigate the organization of talk in groups or pairs, the turn might be an appropriate unit for analysis. As with all studies, classroom researchers should be careful not to reduce the data too far. For example, in a study of the quantity of student talk in the classroom, segmenting the data at the word level would not distin-

guish between learners who have a few extended turns and learners who have many short terms. Data can always be collapsed for analysis (see also chap. 8), but in data collection and coding it is best to keep the categories as narrow as possible.

7.4.2.5. *Considering the Instructional Setting*

Finally, there are setting factors that must be kept in mind when conducting classroom-based research—including whether the research takes place in a foreign or a second language setting, and the particular type of foreign or second language classroom of interest. In the foreign language setting, for example, instruction and expectations for learning might be very different between an EFL course at the university level and an English immersion primary school. It is thus important for researchers to consider the particular characteristics of the context in which they conduct research. To clarify further, if you are a native speaker of English who is teaching or researching English instruction in a foreign language environment with students who do not encounter the language outside of the classroom—you may have difficulty gaining access to sources of information, such as test scores, interviews, and even learners, and your research plans may not fit in with the other instructors' schedules. There is also the issue of subject mortality, or the dropout rate. In some instructional settings it may be acceptable to offer compensation or some kind of reward, whereas this may be inappropriate in other contexts or situations. (Compensation for participation in research was discussed at greater length in chap. 2.)

However, such impediments to conducting research are neither limited nor applicable in blanket fashion to particular language learning situations; even if you are carrying out research in your home country and examining second language learners rather than foreign language learners, you may still find that your research goals and plans are incompatible with those of the instructors with whom you need to work. In addition, in a second language context, it can be more difficult to control all of the variables in quasi-experimental research. For example, the students will be exposed to the target language outside their schools or language programs, thus making firm conclusions about the effect of any treatment more questionable, and perhaps requiring more wide-ranging collection of data by classroom ethnographers. In any case, it is clear that foreign and second language learning research situations must be considered with regard to their own particular characteristics.

7.4.2.6. Summary of Problematics

In pointing out the concerns listed previously, we do not wish to discourage novice or experienced researchers from investigating areas of interest in second language classrooms, but rather to emphasize that classroom research is a particularly complex and multifaceted endeavor that must be planned carefully. We must also stress the importance of flexibility at this point. Even the most carefully designed studies rarely go exactly according to plan in second language classrooms; unforeseen events and problems arise from many sources, and matters that might otherwise be trivial can require the use of quick thinking and adaptation—from there being an odd number of students in the classroom when an experiment calls for pair work to some students having to leave early and not being able to complete the tasks. However, if researchers are aware of this likelihood in advance and can be patient, flexible, and ready to utilize alternate contingency plans, classroom research is ultimately extremely valuable for the field of second language research.

7.5. PURPOSES AND TYPES OF RESEARCH CONDUCTED IN CLASSROOM SETTINGS

Despite the concerns that need to be addressed when carrying out classroom-based research as discussed previously, many successful studies have been carried out in a wide range of second language and foreign language classroom contexts. Some of the different types of research have included large-scale investigations of the effects of instruction, smaller-scale analyses of activities or lessons in classroom settings, detailed ethnographies of specific classes over time, research on learners' aptitude and learning strategies, and qualitatively oriented descriptions of classroom discourse. A range of topics has also been investigated through action research in classrooms, usually carried out by individual instructors on what works in their own instructional contexts. Helpful summaries of the many different classroom research studies can be found in Ellis (1990, 1994). Next, we briefly illustrate methodological practices in two quite different types of research carried out in classroom settings. Despite the wide range of existing classroom research, here we focus narrowly on two different types for illustrative purposes. First, we address traditional classroom-based research by focusing on work carried out on the role of instruction in second language learning. Then, we move on to describe a different type of research set in classrooms, known as action research.

7.5.1. the Relationship Between Instruction and Learning in Second Language Classrooms

The role of instruction in the acquisition of a second language has been studied using a number of different methods in second language classrooms. The majority of researchers have examined the role of instruction in the context of a particular second language learning theory, with a second concern being to inform pedagogical practices.

For example, VanPatten and Cadierno's (1993) study examined the relationship between explicit instruction and input processing (i.e., perceiving the relationship between grammatical form and meaning). In a study using intact classes, VanPatten and Cadierno compared three groups of learners of Spanish as a second language. One class received traditional instruction in object pronoun placement, the second received "processing" instruction on the same topic, and the third received no instruction on this grammar point. In the traditional instruction group, students received explanations on the form and position of direct object pronouns in Spanish and completed typical oral and written classroom exercises, including production exercises. Processing instruction, on the other hand, involved contrasting the forms, presenting the object pronouns, and explaining important points on pronoun position in Spanish. These learners participated in reading and listening exercises in which the focus was the comprehension of object pronouns. Comparing pretest and posttest results for the three groups, the researchers found that the processing instruction group significantly outperformed the other two groups, leading VanPatten and Cadierno to conclude that "instruction is apparently more beneficial when it is directed toward how learners perceive and process input rather than when instruction is focused on having learners practice the language via output" (1993, p. 54). This study is typical of many carried out in the instructed second language acquisition paradigm, particularly in the area of input processing, in which intact classes, different instructional treatments, and a series of pre- and posttests are often utilized.

Other researchers have investigated the role of instruction in second language development by examining different approaches to instruction. For example, there is a great deal of work on the utility of the focus on form approach to instruction. Such research has tended to center on questions of how best to draw learners' attention to form, when to focus on form, and whether one type of focus on form is more effective than another (Spada, 1997). Like the processing instruction study described earlier, research on focus on form has also often involved intact classes receiving different in-

structional methods. For example, Williams and Evans (1998) addressed the question of whether some linguistic forms were more amenable to focus-on-form instruction than others. They used three intact classes: a control group, a group that received an input flood of positive evidence, and a group that received explicit instruction and feedback. For the two experimental groups, the treatments were rotated for the different linguistic forms under investigation. That is, each class received different treatments for participial adjectives and passives. However, the researchers ensured that the instructional materials were not only appropriate for and similar to, but also integrated with the normal activities and focus of the course. Each treatment lasted for approximately 2 weeks. The analysis, which combined quantitative and qualitative data, suggested that not all forms were equal in terms of the effectiveness of focus-on-form activities, and that individual learners could vary greatly in terms of readiness and ability to learn.

Other research on focus on form involving a different methodological approach has been carried out by Ellis, Basturkmen, and Loewen (2001). Ellis et al. recorded and examined a large database of naturalistic classrooms in a descriptive study of what practicing teachers do, finding both teacher-generated and learner-generated incidental focus on form in meaning-based ESL lessons, and reporting that preemptive focus-on-form techniques occurred as frequently as did reactive techniques (38% student initiated, 10% teacher initiated). In a similar series of descriptive studies, Lyster (1998a, 1998b; Lyster & Ranta, 1997) also examined the different techniques that teachers used when reacting to student errors, suggesting that some types of feedback facilitate student responses more than do others.

Research has also been conducted on the effectiveness of instruction for younger learners. For example, in a well-known series of studies involving many years of collaboration with classroom teachers, Lightbown, Spada, and their colleagues described the ESL development of young Francophone learners in Canada, using both description and experimentation to investigate the roles of instruction and error correction. Spada and Lightbown (1993) examined the impact of instruction on question formation in ESL. Following a 2-week period of explicit instruction and corrective feedback, they found that learners improved and maintained their gains on a delayed posttest conducted 5 weeks later. Illustrating the many complexities involved in second language classroom research, Spada and Lightbown reported that control group comparisons were not possible because their control group teacher had used instruction and correction techniques similar to those of the experimental group teachers, despite the researchers' as-

sumptions (based on several data points) that her focus was to be meaningful only in a communicative approach.

As discussed earlier, research on the effects of instruction on second language learning has involved a number of different types of methodologies, ranging from observational to experimental with combined methods and different levels of teacher involvement in the process. Findings have yielded mixed results, with some studies indicating that instruction promotes learning and others suggesting that instruction has little effect. There are many possible reasons for the disparity in findings. As shown previously, it can be more difficult in classroom than in laboratory studies to isolate variables for study, opening possibilities for intervening variables to influence the research findings. Additionally, classrooms vary in significant ways, and instructional practices that seem to enhance learning in one setting may not do so in a different setting. And, of course, the different types of methodologies employed and the different units of analysis and measures of learning all contribute to the difficulty in comparing results.

In an attempt to determine whether there is an overall pattern of positive effects of instruction, Norris and Ortega (2000) performed a meta-analysis of studies of classroom instruction. (Meta-analyses are discussed in more detail in chap. 9 on statistics.) Simply put, meta-analyses examine the findings of a range of different studies and try to synthesize them. Norris and Ortega's overview suggests that instruction does promote second language learning. While lending support to more explicit instructional approaches, their important analysis also illuminates the necessity of considering the nature of the classroom setting, the instructional style, and the many intervening variables when carrying out and interpreting research set in second language classrooms. In their words:

A more complex agenda has begun to unfold within L2 type-of-instruction research that investigates not only the relative effectiveness of particular instructional techniques but also the potential impact of a range of moderator variables (e.g., learner factors such as aptitude, age, and learning style; linguistic factors, such as the relative structural complexity of L2 forms; cognitive factors, such as the learner developmental readiness, degree of noticing; and pedagogical factors, such as timing, duration and intensity of instruction, and integration of interventions within the language curriculum) [R]esearchers will need to turn to more rigorous practices for experimental and quasi-experimental designs. (p. 502)

7.5.2. Action Research

7.5.2.1. Definitions

Although there is little general agreement as to an all-encompassing definition of action research, it is important to realize that action research can be defined and is being implemented in many different ways in the field. For example, Wallace (1998) maintained that action research is “basically a way of reflecting on your teaching ... by systematically collecting data on your everyday practice and analyzing it in order to come to some decisions about what your future practice should be” (p. 4). In this view, action research is a mode of inquiry undertaken by teachers and is more oriented to instructor and learner development than it is to theory building, although it can be used for the latter. Although according to Chaudron (2000), action research does not “imply any particular theory or consistent methodology of research” (p. 4), several steps in the action research process have been usefully identified by action researchers. For example, Nunan (1993) provided a helpful overview of the process involved in conducting action research. In all empirical research on second language classrooms—whether effect-of-instruction, descriptive, or action research—the investigators share similar goals. These include wanting a better understanding of how second languages are learned and taught, together with a commitment to improving the conditions, efficiency, and ease of learning.

7.5.2.2. Theory and Background to Action Research

Teachers can bring a wealth of background knowledge and experience to the research process, offering a unique perspective on the dynamics of second language learning and teaching. Also, teachers may believe that others’ research findings are not sufficiently related or applicable to their own unique teaching situations (Crookes, 1993). As Johnson (1992) noted, when discussing research initiated and carried out by teachers, “if what is missing from the research on classroom language learning is the voices of teachers themselves, then the movement provides ways for teachers’ voices to be heard and valued” (p. 216). Action research is one form of teacher-initiated research. Crookes (1993) provided a useful discussion of the origin of the term, suggesting that “in action research it is accepted that research questions should emerge from a teacher’s own immediate concerns and problems” (p. 130). In contrast to most second language classroom research that is carried out by parties outside the classroom for the purposes of theory

construction and testing, action research is typically conducted by practitioners in order to address an immediate classroom problem or need (Allwright & Bailey, 1991). Like most research, action research usually stems from a question or problem, involves gathering data, and is followed by analysis and interpretation of those data and possibly a solution to the research problem. This can be followed by communication of the findings to others and sometimes by a change or modification to current practice.

7.5.2.3. Action Research in Practice

Before we outline some of the common steps involved in carrying out action research, it is important to note that not all action researchers agree on a process for doing action research, any more than they agree on the nature and content, or even the appropriate name for such research, which is sometimes referred to as “collaborative research” or “practitioner research” or “teacher research.” For example, McDonough and McDonough (1997), with reference to “researcher-generated” and “teacher-initiated” research, discussed the potential tension inherent in referring to teaching as “action” and research as “understanding,” pointing out that both teachers and researchers can do both types and both be parties in research. Allwright and Bailey (1991) also referred to the dynamic nature of the action research framework, suggesting that all research centered on the classroom can be viewed under the unifying characteristic of attempting to understand what goes on in the classroom setting. With this in mind, we now turn to a discussion of the practice of action research.

First, practitioners identify problems or concerns within their own classrooms. For example, a practitioner may be concerned that the students seem to have particular problems with writing an essay. Next, the practitioner may conduct a preliminary investigation in order to gather information about what is happening in the classroom; for instance, the instructor may carefully observe the students during writing classes, examine their written products, and note where problems seem to arise. In this data-gathering phase, the practitioner may decide to create a database with information gathered from multiple sources. As discussed in chapter 6 on qualitative research, triangulation—or the process of obtaining data from more than one source—is an important factor in many types of research, including action research. For our second language writing example, the practitioner may decide to supplement the information gathered from classroom observations and analyses of the students’ written work with other sources of information, such as discussions with colleagues, questionnaires or diary

entries tapping the students' perspectives, verbal protocols or think-alouds produced by the students while writing, and/or the administration of writing tests (e.g., timed essays) in order to gather more information on the students' strengths and weaknesses.

Based on the information obtained in the data, or sometimes before the data are collected, the practitioner may form assumptions or hypotheses. For example, if some of the students seem to have trouble finding and using low-frequency vocabulary words that are required to discuss the topic of the essay, the instructor may then devise and implement some form of intervention or treatment to address that problem. This could constitute a range of techniques, including a new method of teaching vocabulary, or a new technique for raising students' awareness of their own problems and providing them with resources to solve them. Finally, the instructor might evaluate the effects of this practice. This could be done, for example, through another round of data gathering, in which the instructor uses such techniques as observations, questionnaires, verbal protocols, or tests, or simply asks the students for their perspectives. These approaches might help the instructor to determine whether or not the students have benefited from the treatment and also to ascertain the learners' own views about the change in instructional practice, including if and how they feel they have benefited. If the outcome on essay writing is positive, the practitioner may disseminate the results of the process at this point, or return to the stage of reflection. In disseminating results, it is important to remember that much action research is not intended to be generalized. It is situated, or context dependent. In cases in which instructors' treatments, changes in practice, or actions have not been effective, they can consider what other measures could be taken to improve the students' writing. If the changes have been effective, they can consider what else could be done to further support their writing efforts. As can be seen, action research of this form is a cyclic process, and one that many teachers engage in as part of their everyday practice.

Action research is often motivated by teachers' curiosity and their wish to understand their classrooms. An example of this is the following study of the effectiveness of introducing second language adults to reading in English, carried out by Tse (1996). Tse implemented a reading instructional program in her class of adult ESL learners. During the reading program, the learners, none of whom had ever previously read a book in English, read six novels. The class also participated in activities and discussions based on the ideas the stories introduced and kept regular reports of their experiences with the

reading. Tse found that the learners' orientation to reading and reading behaviors changed throughout the study. The learners' attitudes toward reading in English became increasingly positive. Additionally, learners reported that they relied less on dictionaries and were better able to focus on the comprehension of the text as a whole as the semester progressed. This study demonstrates a typical use of teacher research: The teacher identified a question she wanted to investigate in her classroom, and then gathered and analyzed data from her class to determine how well the instruction worked.

As with research on instructed second language acquisition, concerns have also been expressed about various kinds of action research. For example, many types of action research do not typically utilize control groups, and it is often easy to lose sight of concerns with validity or reliability. A further question crops up about how to resolve potential conflicts that arise when the intuitions of teachers run counter to empirical findings about second language learning. It is hard to know exactly how to deal with these criticisms, especially in the light of some discussions (e.g., Johnson, 1992) that action research might best be considered as an independent genre with its own features and standards, and a legitimate rejection of quantitative paradigms. Essentially, it may not always be appropriate to hold action research, or other evolving research paradigms, to the same standards as more established research. However, if action research is intended to inform a wide research community, it will need to meet the basic standards for publication and presentation. Conducted in the complex, dynamic context of the classroom, action research can be "difficult, messy, problematic, and, in some cases, inconclusive" (Nunan, 1993, p. 46). Nevertheless, action research can provide valuable insights both to individual teachers and to the field of second language learning.

7.6. CONCLUSION

As noted earlier in this chapter, second language learning theory is unlikely to be fully developed without some understanding of how second languages are learned in the classroom and, consequently, how they may be more effectively taught. Second language classroom research, regardless of the specific approach taken, allows researchers and teachers to better understand the multitude of factors involved in instruction and learning in different contexts, enhancing our insights into how languages are learned and should be taught. It is worthwhile to note that in recent years, together with

the general trend toward the use of multiple methods in classroom research methodology, collaborative approaches to research are becoming increasingly common and valued, with language teachers and researchers working together as a team to investigate various aspects of second language learning. In the following chapter we deal with coding of data in a range of contexts.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Examine the COLT observation scheme discussed in this chapter. Devise a research question and a short description of a possible research context for a study using the COLT.
2. Consider the context you wrote about for question 1. How could you triangulate the data in this study? How might introspective methods (e.g., uptake sheets, stimulated recall, or diaries) be used to better understand the research phenomenon?
3. Reflect on the following generalized research question: "How are idioms acquired by college-aged ESL learners?" How would you go about setting up a classroom research study? What considerations would you need to make (e.g., recordings, consent, etc.)? Now consider a similar study conducted in an elementary school EFL classroom in Japan. How would you change the study? What different issues might arise in this context?
4. Some researchers (e.g., Foster, 1998) have claimed that research findings from laboratory contexts cannot be applied to classroom settings. Do you agree or disagree with this position? Why?
5. Teachers may choose to carry out action research in their own classrooms to find ways to improve their own teaching. If you are currently a classroom teacher, write a list of questions about your own classes that you would be interested in researching. Try to write at least five questions, noting why they might be important for your teaching. If you are not a current teacher, consider a language learning class in which you previously taught or participated, and come up with five questions that you would be interested in investigating.

Coding

Once data are collected, it is necessary to organize them into a manageable, easily understandable, and analyzable base of information. In this chapter, we discuss some of the main ways of accomplishing this task. We provide an overview of the various processes involved in data coding, including transcription and the preparation of raw data for coding, the modification or creation of appropriate coding systems (depending on the type of data and the research questions), the issue of reliability measures for coding, and the implementation of coding. We also present examples of some common models and custom-made coding systems, taking both quantitative and qualitative research concerns into account. Finally, we discuss questions related to how and when to decide how much and what to code.

8.1. PREPARING DATA FOR CODING

Some types of data can be considered ready for analysis immediately after collection; for example, language test scores such as those from the TOEFL. However, for other types of data, after they are collected they need to be prepared for coding. This chapter focuses primarily on coding of natural data. Coding involves making decisions about how to classify or categorize particular pieces or parts of data. It is helpful to bear in mind Orwin's (1994) comment when preparing to code data: "Coding represents an attempt to reduce a complex, messy, context-laden and quantification resistant reality to a matrix of numbers" (p. 140).

There is a wide range of different types of data in second language research. For example, raw data may be oral and recorded onto audio and/or videotapes; they may be written, in the form of essays, test scores, diaries, or even checkmarks on observation schemes; they may appear in electronic format, such as responses to a computer-assisted accent modification pro-

gram; or they may be visual, in the form of eye movements made while reading text at a computer or gestures made by a teacher in a classroom. They may include learners talking to each other, to native speakers, to teachers, or to themselves in monologues. In short, it is important to recognize that a wide variety of data can be collected for L2 studies.

One common type of second language data is oral. Oral data may come from a range of sources, including, for example, native speaker–learner interviews, learners in pairs carrying out communicative tasks in a laboratory setting, or learners in small groups and their teacher in a noisy L2 classroom setting. Oral data usually need to be transcribed in some way for coding and analysis.

8.1.1. Transcribing Oral Data

8.1.1.1. Transcription Conventions

The process of transcription varies depending on the research goals. As is discussed in more detail later in this chapter, it is not always the case that every utterance of each learner (and/or teacher or native speaker) on a tape will need to be transcribed. In some cases, only the features of interest for the study are transcribed. In other cases, researchers may decide it is sufficient simply to listen to the data and mark on a coding sheet or schedule whether features are present or absent. Either way, interesting examples and exceptions to patterns are usually transcribed for later use in illustrating trends.

Depending on the level of detail required, the time necessary for transcription can vary dramatically. In cases where partial transcriptions are made, the process can proceed quite quickly, taking only about 1 or 2 hours of researcher time per hour of data. However, in other cases—such as the careful and detailed transcription required for conversation analysis of second language data—countless minute aspects of the conversation must be transcribed, leading to as much as 20 hours of researcher time for the transcription of 1 hour of straightforward dyadic conversation and up to 40 hours for a 1-hour recording of overlapping small-group work and conversation (Markee, 2000). Transcriptions are often referred to as broad, including less detail, or narrow—meaning that they are very finely detailed. Transcriptions can be made more easily in second language research by utilizing two tools. The first is an appropriate set of transcription conventions, and the second is a transcription machine.

Simply put, transcription conventions are used to facilitate the representation of oral data in a written format. Conventions can be useful both

for coding and for providing examples from the data when writing up the results of the research. Although there are no generally agreed-on conventions common to all studies, researchers may recognize certain symbols; for instance, the use of dots (ellipses) to convey pauses or silence is quite common. Transcription conventions should match the object of inquiry in the study. For example, if emphasis (or stress) is being investigated as part of a teacher's feedback techniques in the classroom, it will be important to mark emphasis very transparently and distinctly in the transcription and the coding system. Some researchers use boldface type for this purpose, as in "You **have** the ball in your picture." Others might put the emphasized word in all capitals, as in "You **HAVE** the ball," or they might underline it, as in "You have the ball." It might also be necessary for the study to judge and mark the degree of emphasis; for example, a very strongly emphasized item might be both boldfaced and underlined, as in "You **have** the ball," whereas a less strongly emphasized word would be either bolded or underlined. Regardless of the system chosen, the conventions should be explained at the end of the transcripts. For instance, in their study of scaffolding in L2 peer revision, De Guerrero and Villamil (2000) pointed out that they used the following notations to convey different meanings in their transcripts:

- *italics* Italics are employed to cite a letter, word, or phrase as a linguistic example, including Spanish words
- [brackets] Brackets enclose actual Spanish words said by students
- (parentheses) Explanation by authors
- a sequence of dots ... A sequence of dots indicates a pause
- **boldface** Words were said in English (text which is not in English was said in Spanish)
- "quotation marks" Quotation marks indicate participants are reading from the text (De Guerrero & Villamil, 2000, p. 56)

An example of transcription conventions that provide guidelines and notation for different levels of detail appears in Table 8.1. Appendixes H and I provide two other examples of transcription conventions, including one developed specifically for use in second language classrooms.

TABLE 8.1
Sample Transcription Conventions

Spelling: Normal spelling is used for the NNSs and, with a few exceptions (“y’d” for “you’d”; “c’n” for “can”) for the NS.

Intonation/Punctuation: Utterances do not begin with capital letters; normal punctuation conventions are not followed; instead, intonation (usually at the end of a clause or a phrase) is indicated as follows:

At the end of a word, phrase, or clause

- ? Rising intonation
- Falling intonation
- , "Nonfinal intonation" (usually a slight rise)

No punctuation at clause end indicates transcriber uncertainty

Other:

- (?) or () Incomprehensible word or phrase
- (all right) A word or phrase within parentheses indicates that the transcriber is not certain that s/he has heard the word or phrase correctly
- [Indicates overlapping speech; it begins at the point at which the overlap occurs
- = Means that the utterance on one line continues without pause where the next = sign picks it up (latches)
- y- A hyphen after an initial sound indicates a false start
- (.) A dot within parentheses indicates a brief pause
- ((laugh)) Nonlinguistic occurrences such as laughter, sighs, that are not essential to the analysis are enclosed within double parentheses
- CAPITALS Capital letters are used for nonverbal information important to the analysis (e.g., nods, gestures, shifts in posture or position)
- LH Left hand
- RH Right hand
- NOD Refers to one nod
- NODS Refers to more than one nod
- NODS— Refers to nodding accompanying speech, with hyphens indicating how long the nodding (or other behavior) continues
- HS Refers to one head shake
- HSs Refers to more than one head shake;
- HSs— Refers to head shakes accompanying speech, with hyphens indicating how long the head shaking continues

Note: If a nod or head shake does not accompany speech, it is indicated before or after the speech that it precedes or follows; if it accompanies speech, it is represented on a separate line beneath the speech it accompanies. Other nonverbal information is positioned below the speech with which it co-occurs. Gass, S., & Houck, N. (1999). *Interlanguage Refusals* (p. 209). Berlin: Mouton de Gruyter. Copyright © 1999 by Mouton de Gruyter. Reprinted with permission.

8.1.1.2. *Transcription Machines*

Transcription machines make the process of transcribing data significantly easier. Transcription machines usually have a foot pedal so that both hands are free for typing; headphones so that others in the room are not disturbed and so that the transcriber is not disturbed by other distractions; and controls that can be used to adjust the rate of the speech, make typing proceed more quickly, and make it easier to distinguish individual voices. Transcription machines usually have a feature that allows researchers to rewind tapes automatically by a set number of seconds in order to check what they have heard and typed. These machines can be purchased for various types and sizes of cassettes.

8.1.1.3. *Technology and Transcription*

Technology is also changing the process and product of transcriptions. Digital recording equipment is becoming more reasonably priced and accessible, and online controls and software for playback can be customized to make transcription of digital data easier. For cases in which native speaker data need to be transcribed, automatic speech recognition software is improving and, with digital data, could eventually automate the bulk of the transcription task. However, at this time, speech recognition technology does not handle nonnative accents very well regardless of the language of input. Another way in which technology is changing transcription is the increasing use of online journals in which text can be easily and inexpensively manipulated. For example, different colors can be used to represent different speakers or overlap, and multimedia (e.g., short audio or video clips in which readers can actually hear what a learner has said while reading the transcript) can be presented together with the native-speaker or teacher prompts—assuming, of course, that the appropriate permissions have been obtained from all individuals whose voices or images appear in the multimedia clips.

8.2. DATA CODING

Transcriptions of oral data can yield rich and extensive second language data, but in order to make sense of them they must be coded in a principled manner. Data coding, simply defined, entails looking for and marking patterns in data regardless of modality. In this section, we discuss some of the standard ways to present and summarize data, together with some exam-

ples of coding schemes. It is important to note that there is a range of different types of measurement scales that a researcher might employ in second language research. Naturally, the way the data are coded depends in part on the scales used to measure the variables. As discussed in chapter 4, these scales include nominal (often used for classifying categorical data, e.g., nationality, gender, and first language), ordinal (often used for ranking data, e.g., proficiency scores), and interval scales (often used for simultaneously ranking data and indicating the distance, or intervals, between data points).

8.2.1. Coding Nominal Data

Nominal data include cases in which “entities may be the same or different but not ‘more’ or ‘less’... as an example, the part of speech of a given word in a particular sentence, or interpretation of a sentence, is a nominal variable: a word either can be classified as an adjective or it cannot” (Butler, 1985, p. 11). In general, there are two ways nominal data can be coded, depending on whether the research involves a dichotomous variable (i.e., a variable with only two values, e.g., +/– native speaker) or a variable with several values. When dealing with dichotomous variables, researchers may choose to employ signs such as + or –. Alternatively, and particularly when working with a computer-based statistical program such as SPSS or SAS (see chap. 9), researchers may wish to use numerical values (e.g., 1 and 2). For example, in the small database illustrated in Table 8.2, the data have

TABLE 8.2
Sample Nominal Coding: Dichotomous Variables

<i>Code for Participant Identity</i>	<i>Native Speaker Status (1 = native speaker, 2 = nonnative speaker)</i>
A. B.	1
C. P.	2
D. U.	2
Y. O.	2
J. K.	2
H. A.	1
M. B.	1

been coded with numbers to show which participants are native speakers of English, the target language for this hypothetical study. The number 1 is used to indicate native speakers, and the number 2 is used to indicate speakers whose native language is not English.

If the data are not dichotomous and the researcher has to deal with a variable with several values, additional numbers can be used to represent membership in particular categories. For instance, to code the native languages of each of these fictional study participants, a numerical value could be assigned to each of the languages spoken (e.g., Arabic = 1, English = 2, German = 3, Spanish = 4, etc.), as in Table 8.3, which shows, for example, that M. B.'s native language is German.

8.2.2. Coding Ordinal Data

Ordinal data are usually coded in terms of a ranking. For example, with a dataset consisting of test scores from a group of 100 students, one way to code these data would be to rank them in terms of highest to lowest scores. The student with the highest score would be ranked 1, whereas the student with the lowest would be ranked 100. In this scenario, when multiple students have identical scores, ranks are typically split. For example, if two learners each received the fourth highest score on the test, they would both be ranked as 3.5.

Alternatively, instead of using a 100-item list, the scores could be divided into groups (e.g., the top 25%) and each group assigned a number. For ex-

TABLE 8.3
Sample Nominal Coding: Nondichotomous Variables

<i>Code for Participant Identity</i>	<i>Native Language</i>
A. B.	2
C. P.	4
D. U.	4
Y. O.	1
J. K.	3
H. A.	1
M. B.	3

ample, in the database in Table 8.4, a 1 would signify that the individual scored within the top 25%, whereas a 4 would show that the participant scored in the bottom 25%. In our hypothetical study, M. B.'s test score was in the bottom 25% of learners, and he was ranked 100th.

Dividing learners into ranked groups can be particularly useful when using a test where the researcher does not have full confidence in the fine details of the scoring. For instance, a researcher may not believe that a student who scores 88 is very much "better" than a student who scores only 80. In this case, an ordinal scale could be the appropriate way of indicating that the two students are close together, and better than the other groups, without making claims about differences between those students. Ordinal scales can also be used to roughly separate learners from each other; for example, in a study using a battery of L2 working memory tests, the researcher might be interested in examining the data from learners with high and low working memory scores more closely, but might wish to discount the data from learners in the middle-range scores on the basis that they are not differentiated clearly enough. In this case, the middle 50 percent of learners from Table 8.4 could be assigned as "middle" scorers, and only data from students in the top and bottom 25% would be used. There could also be several other cut-off points besides the exact test scores used for the ranking, including points based on measures of central tendency (discussed in chap. 9).

TABLE 8.4
Sample Ordinal Coding

<i>Student Rank</i>	<i>Code for Participant</i>	
	<i>Identity:</i>	<i>Rank Group</i>
1	A. B.	1
2	C. P.	1
3	D. U.	1
30	Y. O.	2
67	J. K.	3
99	H. A.	4
100	M. B.	4

8.2.3. Coding Interval Data

Interval scales, like ordinal scales, also represent a rank ordering. However, in addition, they show the interval, or distance, between points in the ranking. Thus, instead of simply ordering the scores of the test, we could present the actual scores in a table. This would allow us to see not only which scores are higher or lower (as in the ordinal scale), but also the degree to which they differ. For example, in Table 8.5, participant M. B. scored 4 points on the test and was ranked 100th (in last place) of the students who took the test.

Other data that are typically coded in this way include age, number of years of schooling, and number of years of language study. It should be kept in mind, however, that the impact on learning may be different at different intervals. For example, the difference between scores 1 and 10 may have the same interval as those between 90 and 100 on a test, but the impact is quite different. Similarly, the difference between 2 and 3 years of instruction may be the same interval as the difference between 9 and 10 years. In each case, the difference is only 1 year, but that year might be very different in terms of the impact on language production for a learner who is at the advanced, near-native stage, as compared to a learner who is in the early stages of acquisition. These are issues that merit careful consideration in the coding stages of a research study.

TABLE 8.5
Sample Interval Coding

<i>Student Rank</i>	<i>Code for Participant Identity</i>	<i>Test Score (X/100)</i>
1	A. B.	98
2	C. P.	96
3	D. U.	95
30	Y. O.	68
67	J. K.	42
99	H. A.	8
100	M. B.	4

8.3. CODING SYSTEMS

The measures presented so far represent some of the common first steps in coding data. Bernard (1995) suggested that “the general principle in research is always use the highest level of measurement that you can” (p. 32). By this he meant, for example, if you wish to know about the amount of prior instruction in a particular language, you should ask the learners a question such as “How many years of prior instruction in X-language have you had?” rather than “Have you had 1–2 years, 2–5 years, or more than 5 years of prior instruction?” Basically, if researchers code data using as finely grained a measurement as possible, the data can always be collapsed into a broader level of coding later if necessary, but finely grained categories are harder, if not impossible, to reconstruct after the data are coded. Another way to put this is that, in coding, the categories should always be as narrow as possible. For example, in a study in which “interactional feedback” is to be coded, both recasts and negotiation could be considered as feedback. However, it would be much more judicious to code them separately at first and later decide that these two categories could be collapsed into one “feedback” category than it would be to code them both into one “feedback” category and later decide the research question needed to be addressed by separating them, thus necessitating a recoding of the data.

A range of different coding practices can be used with second language data to allow researchers to gain a deeper understanding of the information they have collected. Usually in the coding process, patterns in the data are indicated in separate records as one examines the data. However, coding is sometimes recorded directly onto the data source, as in the case of interview transcripts or essays, for example. Coding systems are often referred to as sheets, charts, techniques, schemes, and so on. In any case, they should be as clear and as straightforward to use as possible, as we discuss later in this chapter. Many researchers develop a coding scheme based on their specific research questions (unless they are carrying out a replication study, in which case they usually use coding instruments identical to those of the original study). In the second language research field, it would be helpful if researchers made more use of existing coding schemes, because this would facilitate easy comparisons across studies. However, sometimes existing schemes require refinements to capture new knowledge, and sometimes new schemes are required depending on the research question. Coding systems range from those based on standard measures, which have the advantage of increasing the generalizability of the research because they are used by a range of researchers, to highly customized systems developed simply

for the study at hand. Many different schemes have been developed for coding all sorts of second language data, and in this chapter we provide examples of some of them. However, it is important to recognize that it would be impossible to cover the whole range of existing schemes.

8.3.1. Common Coding Systems and Categories

A number of coding units or categories for oral and written data have been proposed over the years. These include such units as the following:

- T-units.
- Suppliance in obligatory context (SOC) counts.
- CHAT convention.
- Turns.
- Utterances.
- Sentences.
- Communication units.
- Tone units.
- Analysis of speech units.
- Idea units.
- Clauses.
- S-nodes per sentence.
- Type-token ratios.
- Targetlike usage counts.

Three of the most common of these—T-units, SOC, and CHAT—are discussed in more detail next, together with information about counting different types and different tokens of coding units.

8.3.1.1. T-Units

A T-unit is generally defined as “one main clause with all subordinate clauses attached to it” (Hunt, 1965, p. 20). They were originally used to measure syntactic development in children’s L1 writing. However, they have become a common measurement in second language research as well, and have served as the basis for several ratio units, such as number of words per T-unit, words per error-free T-unit, and clauses per T-unit. An example of a T-unit is the utterance

After she had eaten, Sally went to the park

This T-unit is error-free; that is, it contains no nontargetlike language. An alternative T-unit

After eat, Peter go to bed

would be coded as a T-unit containing errors. To code using T-units, a researcher may, for example, go through an essay or a transcription and count the total number of T-units; from this number, the researcher could count all the T-units not containing any errors and then present a ratio. For instance, the researcher could say that of 100 T-units used by a learner, 33 contained no errors. T-units have been used as a measure of linguistic complexity, as well as accuracy.

Although commonly employed and sometimes held up as useful because of comparability between studies, the use of T-units has been criticized (e.g., Bardovi-Harlig, 1992; Gaies, 1980). For example, it has been argued that the error-free T-unit measure is not always able to take into account the linguistic complexity of the writing or speech or the severity of the errors (Polio, 1997). In addition, the definitions and types of “error” and the methods of counting errors have varied considerably from one researcher to the next. Nevertheless, T-units remain popular in second language research, in part because they are easy to identify and are relatively low-inference categories.

8.3.1.2. Suppliance in Obligatory Contexts (SOC)

Some second language studies have focused on grammatical accuracy with respect to specified linguistic features. A researcher may be interested in whether a learner has acquired a particular grammatical form such as the simple past tense, the progressive *-ing*, or the third person singular *s*. The learner’s level of acquisition can be measured in terms of how often these features are supplied where they are required. This is commonly known as suppliance in obligatory contexts (SOC). For example, in the sentence “He is singing right now,” the *-ing* is required because this is a context in which the progressive form is obligatory. SOC was first used in early studies of the acquisition of grammatical morphemes by children acquiring English as their first language (e.g., Brown, 1973), but it has also been applied in second language studies.

For instance, SOC was employed as a measure in a 2-year study of 16 adult learners of English reported in Bardovi-Harlig (2000). The focus of the research was the emergence of tense-aspect morphology related to past-time contexts. Two researchers independently coded every verb sup-

plied in a past-time context with respect to its verbal morphology. A computer sorting program was then used to analyze the data, coding the verbs into types and tokens and counting them. Whereas a count of tokens would tally repetitions of the same form as separate instances and might therefore artificially inflate the rates of appropriate use by including multiple occurrences of common verbs such as *was* and *went*, the counting of types would enable the researchers to provide a conservative view of the acquisition of tense-aspect morphology. Thus, to calculate the rates of appropriate use of the past tense in this study, the researchers used the ratio of the number of past-tense forms supplied to the number of obligatory environments, expressing the rates as percentages of appropriate use.

Although SOC is a useful measure of morpheme use in required contexts, Pica's (1984) study of the acquisition of morphemes brings up a common criticism of SOC: namely, that it does not account for learners' use of morphemes in inappropriate contexts. To address this, Pica used target-like usage (TLU) as an additional measure. This takes into account both appropriate and inappropriate contexts.

8.3.1.3. CHAT

Whereas T-units and SOC tend to focus primarily on linguistic accuracy, the CHAT system is aimed at discourse. CHAT was developed as a tool for the study of first and second language acquisition as part of the CHILDES (Child Language Data Exchange System) database (see chapter 3 for more on the CHILDES database). It has become an increasingly common system for the coding of conversational interactions and employs detailed conventions for the marking of such conversational features as interruptions, errors, overlaps, and false starts (MacWhinney, 1999, 2000). A standard but detailed coding scheme such as CHAT is particularly useful in qualitative research. For example, in conversation analysis, researchers typically eschew a focus on quantifying data and concentrate instead on portraying a rich and detailed picture of the interaction, including its sequencing, turn taking, and repair strategies among participants in a conversation (Markee, 2000). Thus, whereas a researcher conducting a quantitative study might code a transcript for errors in past-tense formation, another researcher undertaking conversation analysis might mark the same transcript with a much more detailed coding system, marking units such as length of pauses and silences, stress, lengthening of vowels, overlaps, laughter, and indrawn breaths. It is important to realize that many researchers working in qualitative paradigms, including those working on conversation analysis, have argued that quantification of cod-

ing does not adequately represent their data. For example, Schegloff (1993) pointed out that laughter is responsive and its positioning is reactive to conversational structure, leading him to conclude that a quantified measure such as “laughs per minute” would be inadequate to describe the dynamic nature of laughter during conversation.¹

It is important for researchers to keep in mind that regardless of the potential utility of standard coding systems in increasing the generalizability of findings, the goal is always to ascertain how best to investigate one’s own research questions. In much second language research, preexisting coding systems and categories are the exception rather than the rule. Many researchers develop their own systems. In the next section, we provide examples of custom-made coding systems based on five different research areas: grammatical development (question formation), negative feedback, classroom interaction, L2 writing instruction, and task planning.

8.3.2. Custom-Made Coding Systems

8.3.2.1. Question Formation

An example of a custom-made scheme was used by Mackey and Philp (1998) in their exploration of the relationship between interactional feedback in the form of recasts and the development of question formation by learners of English as a Second Language. They collected data from learners carrying out communicative tasks at four intervals in a pretest/posttest design. Because their focus was on whether or not the learners’ questions developed, the researchers needed a coding scheme that would allow them to identify how the learners’ question formation changed over time. They based their coding system on the custom-made six-stage sequence of questions in the morpho-syntactic sequence adapted for ESL by Pienemann and Johnston (1986). Pienemann and Johnston’s sequence has been used in a wide range of studies, including Spada and Lightbown’s (1993) study of the effects of instruction on question formation and interaction studies carried out by Mackey (1999), Philp (2003), and Silver (1999).

To code the data, Mackey and Philp first designated the questions produced by their child learners as belonging to one of the six stages based on the Pienemann-Johnston hierarchy. A modified version of the stage descriptions used by Mackey and Philp and by a number of other researchers (as noted above) appears in Table 8.6. It is important to note that in coding

¹As discussed, with examples, in chapter 7, a comprehensive set of coding systems has also been developed for classroom observations.

TABLE 8.6

Coding for Questions: Tentative Stages for Question Formation

Stage 1	<i>Single words or sentence fragments</i> One astronaut outside the spaceship?
Stage 2	<i>Canonical word order</i> It's a monster in the right corner? The boys throw the shoe? He have two house in the front?
Stage 3	<i>Wh-fronting and Do-fronting</i> How many planets are in this picture? Where the little children are? What the dog do? What color the dog? Do you have a shoes on your picture? Does in this picture there is four astronauts?
Stage 4	<i>Pseudo inversion</i> a. Inversion in wh-questions with copula Where is the sun? b. Inversion in yes/no questions with auxiliaries other than do The ball is it in the grass or in the sky?
Stage 5	<i>Do-second: Inversion with do in wh-questions</i> How many astronauts do you have? <i>Aux second: inversion with other auxiliaries in wh-questions</i> What's the boy doing?
Stage 6	<i>Question tag</i> You live here, don't you? <i>Negative question</i> Doesn't your wife speak English? <i>Subordinate clause</i> Can you tell me where the station is?

Note: Adapted from the developmental stages described in Pienemann and Johnston (1986). Spada, N., & Lightbown, P. (1993). Instruction and the development of questions in the L2 classrooms. *Studies in Second Language Acquisition*, 15(2), 222. Copyright © 1993 by Cambridge University Press. Reprinted with the permission of Cambridge University Press and with the permission of Nina Spada and Patsy Lightbown.

according to these descriptions, not every question produced by the learners is codable because some questions do not fit into the scheme, given that not all question types are included in the scheme.

Another important point is that grammaticality is not coded using this scheme; because it was designed by Pienemann and his associates in an attempt to capture processing capabilities and linguistic complexity, linguistic accuracy is not the primary focus. The goal of the description of learners' developmental stages is to capture processing capabilities and developing linguistic complexity.

Following the assignment of each question to a particular stage, the next step in carrying out coding based on this hierarchy was to determine the highest level stage that the learners reached. Pienemann and Johnston's (1986) model suggested that learners can be assigned to a stage. Assignment to a stage is generally determined by the use of two different forms. A more conservative version of this criterion was adopted by later research, with the result that Mackey and Philp determined that two productive usages of two different question forms on at least two different tasks were required for learners to be said to have reached a given stage. Thus, the second step of the coding involved the assignment of an overall stage to each learner, based on the two highest-level question forms asked in two different tests. It was then possible to examine whether the learners had improved over time. Table 8.7, based on constructed data, shows the second level of the coding, in which each learner has been assigned to a particular stage.

As can be seen in Table 8.7, learner AB continued throughout the study at the third stage. If learner AB were in a control group, this would generally be an expected outcome. Learner AA began the study at Stage 3 and then continued through the next three posttests at Stage 5. Once this sort of coding has been carried out, the researcher can make decisions about the analysis, such as whether to carry out statistical tests on learning outcomes by comparing each test.

8.3.2.2. Negative Feedback

Oliver (2000) examined whether the provision and use of negative feedback were affected by the age of the learners (adult or child) and the context of the interaction (classroom or pair work). In order to do this, she developed a hierarchical coding system for analysis that first divided all teacher–student and NS–NNS conversations into three parts: the NNS's initial turn, the response given by the teacher or NS partner, and the NNS's reaction. Each part then was subjected to further coding. First, the NNS's

TABLE 8.7
Coding for Question Stage

ID	Pretest				Immediate Posttest				Delayed Posttest			
	Task 1	Task 2	Task 3	Final Stage	Task 1	Task 2	Task 3	Final Stage	Task 1	Task 2	Task 3	Final Stage
AB	3	3	2	3	3	3	3	3	3	3	2	3
AA	3	3	3	3	5	5	4	5	5	5	4	5
AC	3	4	3	3	2	2	3	2	3	3	3	3
AD	3	3	4	3	3	5	5	5	5	3	3	3

initial turn was rated as *correct*, *nontargetlike*, or *incomplete*. Next, the teacher’s/NS’s response was coded as *ignore*, *negative feedback*, or *continue*. Finally, the NNS’s reaction was coded as *respond* (e.g., by incorporating the negative feedback into a subsequent utterance), *ignore*, or *no chance* to react to it. As with many schemes, this one is top-down, sometimes known as hierarchical, and the categories are mutually exclusive, meaning that it is possible to code each piece of data in only one way. Figure 8.1 below represents this scheme graphically.

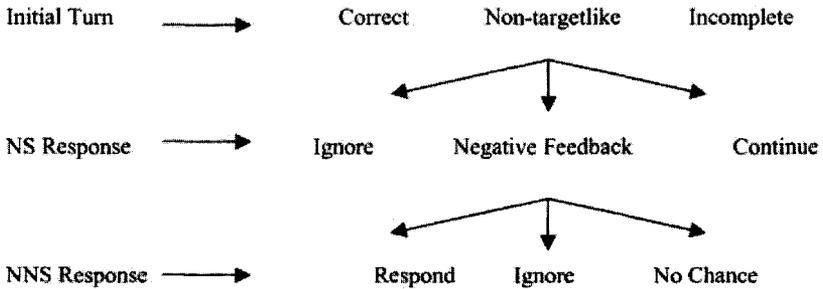


FIG. 8.1. Three-turn coding scheme.

8.3.2.3. Classroom Interaction

Lyster and Ranta (1997) studied the use of linguistic feedback by adolescent students in second language immersion schools. They were particu-

larly interested in which types of feedback were most commonly employed by the teachers and whether certain types of feedback were more likely to be used by the learners, defining use in terms of third-turn uptake. To this end, they examined classroom transcripts for learner errors. When an error occurred, the next turn was examined to determine whether the error was corrected, or whether it was ignored and the topic continued by the teacher or the learner. If the error was corrected, the following turn was examined and coded according to whether the learner produced uptake or whether the topic was continued. Finally, the talk following uptake was examined with regard to whether the uptake was reinforced or the topic continued. The coding categories are illustrated in Fig. 8.2.

8.3.2.4. Second Language Writing Instruction

The following two studies used coding categories that differ from generally form-oriented coding schemes in that the focus was not on the learners' overall knowledge of forms, but on evidence of development following an intervention. Adams (2003) investigated the effects of written error correction on learners' subsequent second language writing. In a pretest/posttest experimental design, university-level Spanish learners wrote short stories based on a set of pictures. During treatment, the learners' stories were rewritten with each form grammatically corrected, and the experimental group was provided with the opportunity to compare their original stories with the reformulated versions. For the posttest, the learners were asked to write their stories again, using the pictures. Adams wanted to determine whether the final versions of the stories would show evidence of learning following the reformulations. The forms in the final essays were coded and compared with those in the pretest as being more targetlike, not more targetlike, or not attempted (avoided).

In a similar study that also examined changes in linguistic accuracy following written feedback, Sachs and Polio (2004) compared three feedback conditions (plus a control group). Having written stories based on pictures, university-level ESL learners in the three experimental groups were provided with the opportunity to compare their original stories with either reformulated or explicitly corrected versions. One of the groups receiving reformulations performed the comparisons while simultaneously producing verbal protocols. A few days later, all four groups were asked to rewrite their stories. Sachs and Polio first coded all of the learners' errors individually. They then segmented the stories and revisions into T-units, examined them side by side, and coded each T-unit as being: at least partially changed

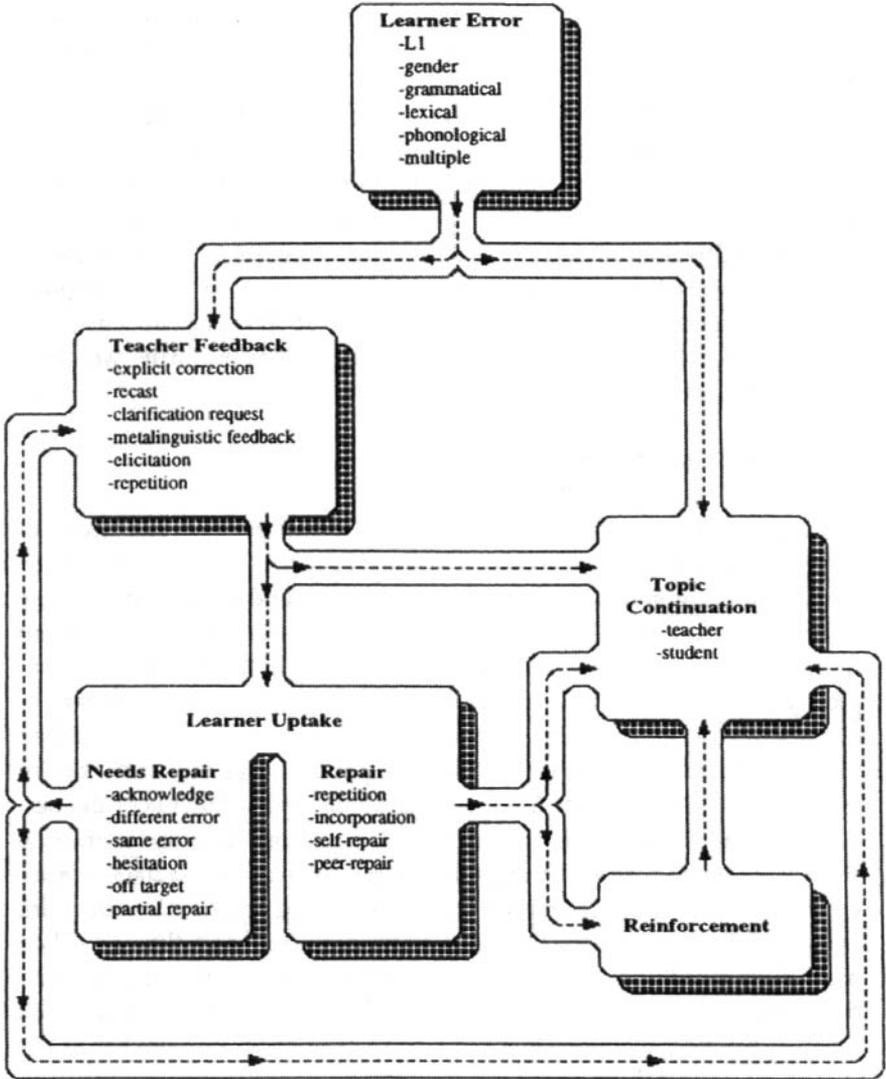


FIG. 8.2. Classroom interaction coding scheme. Source: Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19, 44. Copyright © 1997 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

in a way related to the original error(s) (+), completely corrected (0), completely unchanged (-), or not applicable (n/a) because there had been no errors in the original T-unit.

It is interesting to note that although the studies by Adams (2003) and by Sachs and Polio (2004) focused on changes in linguistic accuracy in L2 writing, the researchers used different coding schemes to fit with the correspondingly different focus of their research questions in order to compare the four feedback conditions with each other. Whereas Adams coded individual forms as more targetlike, not more targetlike, or not attempted, Sachs and Polio considered T-unit codings of "at least partially changed" (+) to be possible evidence of noticing even when the forms were not completely more targetlike.

8.3.2.5. Task Planning

Research on the effects of planning on task performance has often utilized measures of fluency, accuracy, and complexity. Skehan (1998) argued that fluency is often achieved through memorized and integrated language elements. Accuracy is achieved when learners use an interlanguage system of a particular level to produce correct but possibly limited language. Complexity comes about when learners show a willingness to take risks and try out new forms even though they may not be completely accurate. Skehan further claimed that these three aspects of performance are somewhat independent of one another. A range of different measurements of fluency, accuracy, and complexity have been used in the second language literature. In research on tasks and planning for example, Yuan and Ellis (2003) investigated the effects of both pretask and online planning on L2 oral production, using multiple measures of complexity, accuracy, and fluency. They operationalized fluency as (a) number of syllables per minute, and (b) number of meaningful syllables per minute, where repeated or reformulated syllables were not counted. This measure of fluency was chosen because it "takes into account both the amount of speech and the length of pauses" (p. 13). They operationalized complexity as syntactic complexity, the ratio of clauses to t-units; syntactic variety, the total number of different grammatical verb forms used; and mean segmental type-token ratio (this procedure was followed to take into account the effect of text length). They operationalized accuracy as the percentage of error-free clauses, and correct verb forms (the percentage of accurately used verb forms). Their study illustrates the benefits of a coding system that is similar enough to those

used in previous studies that results are comparable, while also finely grained enough to capture new information.

8.3.3. Coding Qualitative Data

Just as with quantitative research, qualitative researchers code data by identifying patterns. However, in qualitative research, coding is usually grounded in the data. In other words, the schemes for qualitative coding generally emerge from the data rather than being decided on and preimposed prior to the data being collected or coded. This process, in which initial categories are based on a first pass through the data, is sometimes known as open coding. Qualitative researchers explore the shape and scope of the emerging categories and investigate potential connections among categories. As more data are coded, researchers also consider aspects such as the range of variation within individual categories. These processes can assist in the procedure of adapting and finalizing the coding system, with the goal of closely reflecting and representing the data.

For example, one way of coding qualitative data can involve examining the data for emergent patterns and themes, by looking for anything pertinent to the research question or problem, also bearing in mind that new insights and observations that are not derived from the research question or literature review may be important. Paraphrases, questions, headings, labels, or overviews can be assigned to chunks of the data. These labels or indicators are usually not precise at the early stages. The data, rather than the theory or framework, should drive the coding. Many researchers try to code the data by reminding themselves that they will need to explain how they arrived at their coding system, keeping track of the data-based origins of each of their insights. Interesting data that are extra to the goals of the study are not discarded; they are kept in mind and possibly also coded. Themes and topics should emerge from the first round of insights into the data, when the researcher begins to consider what chunks of data fit together, and which, if any, are independent categories. Finally, a conceptual schema or organizational system should emerge, by which researchers consider their contribution to the field. At this stage, researchers often ask themselves if they can tell an interesting narrative based on the themes in the data. At this stage they are often ready to talk through their data and the patterns with others, so that input can help them in the stages before they write up their research. This is just one method by which qualitative researchers can code and analyze their data. Denzin and Lincoln (1994) presented a comprehensive picture of the many alternatives.

One problem with developing highly specific coding schemes is that it can be problematic to compare qualitative coding and results across studies and contexts. However, as Watson-Gegeo (1988) pointed out, although it may not be possible to compare coding between settings on a surface level, it may still be possible to do so on an abstract level. Whereas a particular event may not occur in two settings, the same communicative need can exist in both. For example, in examining the relationship between second language learning and attitudes of immigrant children, although one study may focus on the school context and another on the home context, and each may examine different types of events in the data, the overall questions and answers may be comparable.

8.4. INTERRATER RELIABILITY

Regardless of the choice researchers make from the wide range of different types of data coding that are possible, establishing coding reliability is a crucial part of the process. The choice of which coding system to adopt, adapt, or devise ultimately depends on the researcher's goals and the type of study being carried out. However, it is common to ensure that the coding scheme can be used consistently or reliably across multiple coders wherever possible. This is known as interrater reliability, a concept introduced in chapter 4.

Because coding involves making decisions about how to classify or categorize particular pieces of data, if a study employs only one coder and no intracoder reliability measures are reported, the reader's confidence in the conclusions of the study may be undermined. To increase confidence, it is important not only to have more than one rater code the data wherever possible, but also to carefully select and train the raters. It may be desirable to keep coders selectively blind about what part of the data (e.g., pretest or posttest) or for which group (experimental or control) they are coding, in order to reduce the possibility of inadvertent coder biases. In some cases, researchers act as their own raters; however, if, for example, a study involves using a rating scale to evaluate essays from second language writers, researchers may decide to conduct training sessions for other raters in which they explain something about the goals of the study and how to use the scale, provide sample coded essays, and provide opportunities and sample data for the raters to practice rating before they judge the actual data. Another way to increase rater reliability is to schedule coding in rounds or trials to reduce boredom or drift, as recommended by Norris and Ortega (2003). One question that is often raised is how much data should be coded by second or third raters. The usual answer is, as much as is feasible given

the time and resources available for the study. If 100% of the data can be coded by two or more people, the confidence of readers in the reliability of the coding categories will be enhanced, assuming the reliability scores are high. However, researchers should also consider the nature of the coding scheme in determining how much data should be coded by a second rater. With highly objective, low-inference coding schemes, it is possible to establish confidence in rater reliability with as little as 10% of the data. We now turn to a discussion of those scores.

8.4.1. Calculating Interrater Reliability

In addition to training the raters and having as much data as possible scored by more than one rater, it is also crucial to report interrater reliability statistics and to explain the process and reliability estimate used to obtain these statistics.

8.4.1.1. Simple Percentage Agreement

Although there are many ways of calculating interrater reliability, one of the easiest ways is through a simple percentage. This is the ratio of all coding agreements over the total number of coding decisions made by the coders. For example, in Mackey and Oliver's (2002) study of children's ESL development, both researchers and one research assistant coded all of the data. This process yielded an interrater reliability percentage of 98.89%, meaning that there was disagreement over only 1.11% of the data. Simple percentages such as these are easy to calculate and are appropriate for continuous data (i.e., data for which the units can theoretically have any value in their possible range, limited in precision only by our ability to measure them—as opposed to discrete data, whose units might, for example, be limited to integer values). Their drawback is that they have a tendency to ignore the possibility that some of the agreement may have occurred by chance. To correct for this, another calculation is commonly employed—Cohen's kappa (Cohen, 1960).

8.4.1.2. Cohen's Kappa

This statistic represents the average rate of agreement for an entire set of scores, accounting for the frequency of both agreements and disagreements by category. In a dichotomous coding scheme (e.g., coding forms as targetlike or nontargetlike), Cohen's kappa requires that the researcher determine how many forms both raters coded as targetlike, how many were

coded as targetlike by the first rater and as nontargetlike by the second, how many were coded as nontargetlike by the first and as targetlike by the second, and so on. The final calculation of kappa therefore involves more detail on agreement and disagreement than simple percentage systems, and it also accounts for chance.

8.4.1.3. Additional Measures of Reliability

Other measures, such as Pearson's Product Moment or Spearman Rank Correlation Coefficients, may also be used to calculate interrater reliability. These latter two are based on measures of correlation and reflect the degree of association between the ratings provided by two raters. They are further discussed in chapter 9, in which we focus on analysis.

8.4.1.4. Good Practice Guidelines for Interrater Reliability

In most scientific fields, including second language research and associated fields such as education, "there is no well-developed framework for choosing appropriate reliability measures" (Rust & Cooil, 1994, p. 2). Although a detailed examination and comparison of the many different types of interrater reliability measures is beyond the scope of this chapter (for more comprehensive reviews see Carmines & Zeller, 1979; Chaudron, Crookes, & Long, 1988; Gwet, 2001; Pedhazur & Schmelkin, 1991), general good practice guidelines suggest that regardless of which measurement is chosen, researchers should state which measure was used to calculate interrater reliability, what the score was, and, if there is space in the report, briefly explain why that particular measure was chosen. Some researchers also explain how data about which disagreements arose were dealt with; for example, if agreement was eventually reached and the data were included in the analysis, or if data (and how much) were discarded.

There are also no clear guidelines in the field of second language research as to what constitutes an acceptable level of interrater reliability. The choices and decisions clearly have lower stakes than, for example, in the field of medicine. However, the following rough guidelines based on rigorous standards in some of the clinical science research may be of some assistance (Portney & Watkins, 1993):

- For simple percentages, anything above 75% may be considered "good," although percentages over 90% are ideal.

- For Cohen's kappa, 0.81 to 1.00 is considered "excellent." In general, a reader should be concerned with a percentage of less than 80%, because this may indicate that the coding instrument needs revision.

8.4.1.5. How Data Are Selected for Interrater Reliability Tests

As noted earlier, in some second language studies the researchers code all of the data and calculate reliability across 100% of the dataset. However, an alternative is to have the second or third rater code only a portion of the data. For instance, in some studies the researcher may semi-randomly select a portion of the data (say 25%) and have it coded by a second rater (and sometimes by a third or fourth rater as well, depending on the size of the dataset and the resources of the researcher). If this approach is taken, it is usually advisable to create comprehensive datasets for random selection of the 25% from different parts of the main dataset. For example, if a pretest and three posttests are used, data from each of them should be included in the 25%. Likewise, if carrying out an interrater reliability check in an L2 writing study, essays from a range of participants at a range of times in the study should be selected.

It is often necessary to check intrarater reliability as opposed to interrater reliability. In Philp's (2003) study, she coded all of the data. She then recoded 15% of the data herself 6 months later to check for intrarater reliability. Intrarater reliability refers to whether a rater will assign the same score after a set time period. Philp used this system together with a standard check for interrater reliability, also having one third of her treatment transcripts double-coded by six assistants.

8.4.1.6. When to Carry Out Coding Reliability Checks

It is important to realize that if a researcher codes 100% of a dataset him- or herself, and then realizes that the coding system is unreliable, a great deal of unnecessary effort will have been expended, because the coding system may need to be revised and the data recoded. For this reason, many researchers decide to use a sample dataset (perhaps a subset of the data, or data from the pilot test) to train themselves and their other coders, and test out their coding scheme early on in the coding process. Following this initial coding and training, coders may then code the rest of the dataset independently, calculating interrater reliability at the end of the coding process on the data used for the research, rather than for the training.

When space permits, we recommend the following reporting on coding:

- What measure was used.
- The amount of data coded.
- Number of raters employed.
- Rationale for choosing the measurement used.
- Interrater reliability statistics.
- What happened to data about which there was disagreement (e.g., recoded? Not included?).

Complete reporting will help the researcher provide a solid foundation for the claims made in the study, and will also facilitate the process of replicating studies. If a low interrater reliability statistic is reported, this may be an indication that future studies will need to revise the coding system.

8.5. THE MECHANICS OF CODING

After selecting or devising an appropriate coding system, the researcher must determine how to go about coding the data. Implementations of systems vary among researchers according to personal preferences. Some researchers, for example, may prefer a system of using highlighting pens, working directly on transcripts, and marking such things as syntactic errors in one color pen and lexical errors in another, with a tally on each page and a final tally on the first page of the transcript. Other researchers, depending on their particular questions, may decide to listen to tapes or watch videotapes without transcribing everything; they may simply mark coding sheets when the phenomena they are interested in occur, and may decide to transcribe only interesting examples for their discussions. This system may also be used for written data, for which coding sheets are marked directly without marking up the original data. Still other researchers may prefer to use computer programs to code data if their research questions allow it. For example, if a researcher is interested in counting the number of words in different sections of an essay or selecting the central portion of a transcript for analysis, it would be much easier to use a word processor than it would be to do this exercise by hand. If the research questions relate to computer-assisted language learning, many CALL programs automatically record each keystroke a learner makes, and these data can easily be sorted and coded. Likewise, if the researcher wishes to focus on such things as reaction times to certain presentations on a computer screen, or eye movements as learners read and write text, reaction time software would be a possible choice.

8.5.1. How Much to Code?

As suggested previously, not all research questions and coding systems require that an entire dataset be coded. When selecting and discussing the data to code, researchers first need to consider and justify why they are not coding all their data. A second important step is determining how much of the data to code. This process is sometimes known as data sampling or data segmentation. Some researchers may decide it is important to code all of the data, whereas others may decide their questions can be answered by examining a portion of the data. In making decisions about how much and which portions of data to code, another point to consider is that the data to be analyzed must always be representative of the dataset as a whole and should also be appropriate for comparisons if these are being made. For example, if a researcher chooses to code the first 2 minutes of oral data from a communicative task carried out by one group of students and the last 2 minutes from another group of students, the data might not be comparable because the learners could be engaged in different sorts of speech even though the overall task is the same. In the first 2 minutes they might be identifying and negotiating the problem or activity, whereas in the final 2 minutes they might be making choices about how to complete the activity, or even how to communicate the outcomes to others. Another possibility is that the learners may begin to lose interest or feel fatigued by the end. In view of these concerns, the researcher could instead choose to take the middle section of the data—for example, the central 50 exchanges. Whenever possible, if only a portion of the data is being coded, researchers should check that the portion of data is representative of the dataset as whole. Of course, as with everything else about research design, the research questions should ultimately drive the decisions made, and researchers need to specify principled reasons for selecting data to code.

In much of Oliver's work (1998, 2000, 2002), she has made a general practice of coding only the first 100 utterances of each of her extended interactions. For example, in Oliver (2000) she coded the first 100 utterances of each teacher-fronted lesson and of each of the pair-work tasks in her study. Oliver made this decision because, in some cases, the interactions were only a little more than 100 utterances, and she needed a minimum comparable number of units to code. In summary, depending on the research questions and the dataset, a number of different segmentation procedures may be appropriate in second language research.

8.5.2. When to Make Coding Decisions?

Wherever possible, it is best to make decisions concerning how to code and how much to code prior to the data collection process—that is, when planning the study and preparing the protocol. By addressing coding concerns at the beginning—hopefully through a detailed pilot study—the actual collection of data can be fine tuned. For instance, if straightforward coding sheets are designed ahead of time based on research questions and variables, it may become obvious that the proposed data collection procedures cannot provide clear answers to the research questions. This may lead researchers to rework their plans for gathering data so that they can gather more information or different types of information from other sources. The best way to uncover and address such issues is by carrying out an adequate pilot study. This will allow for piloting not only of materials and methods, but also of coding and analysis. Designing coding sheets ahead of data collection and then testing them out in a pilot study is the most effective way to avoid potential problems with the data for the study.

8.6. CONCLUSION

Data coding is one of the most time-consuming and painstaking aspects involved in carrying out a second language research project. There are many decisions to be made, and it is important to remember that many of the processes involved in data coding can be thought through ahead of time and then pilot tested. These include the preparation of raw data for coding, transcription, the modification or creation of appropriate coding systems, and the plan for determining reliability. Careful coding is a key component of good research. In the next chapter we focus on quantitative data and, in particular, on what to do with data once they are coded and ready to be analyzed.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Classify the following as nominal, ordinal, or interval data:
 - a. The number of T-units in a 100-utterance transcript.
 - b. The presence or absence of a verb form in an obligatory context.
 - c. Ten students ranked from 1 to 10 based on their recall of propositions from a reading passage.
 - d. The number of final words in 10 sentences that were recalled correctly in a test of working memory.

- e. The 10 different L1s of a group of 15 ESL children.
 - f. The number of errors in a single student essay.
2. Is it necessary to use transcription conventions? Why or why not?
 3. What are the advantages and disadvantages of using or modifying an existing coding scheme as opposed to devising a scheme of one's own to fit the data?
 4. Once you have a coding scheme, do all your data need to be coded under that same scheme?
 5. How much data should you optimally use for interrater reliability checks on your coding?
 6. Name three methods for interrater reliability checks.
 7. What is intrarater reliability?
 8. What is data segmentation? How do you determine how to segment data?
 9. When is the optimal time to make decisions about coding? What factors should you consider, and what can help you in making your decisions?
 10. Suggest coding categories for the following data. The dataset was collected to address the question of the relationship between frequency of input and the development of third-person singular *s*:
Teacher: She goes to the park in this video.
Learner 1: She go to the park.
Teacher: Goes, she goes there every day.
Learner 2: She went yesterday?
Teacher: She went yesterday, and today she goes again.
Learner 3: Again today she go to the park. She like the park fine.
Teacher: She goes to the park most days. She likes it. She will go tomorrow, she went yesterday, and she goes there today as well.
Learner 3: Always she go to the park. Not too boring for her?

Analyzing Quantitative Data

This chapter presents introductory information about statistics to enable the reader to begin to understand basic concepts. We focus on issues and methods of analysis that are common in second language research. The chapter deals with descriptive as well as analytic measures. It also addresses concepts such as normal distribution, standard scores, and probability, all of which are necessary to an understanding of basic statistical procedures.

9.1. INTRODUCTION

In chapter 8 we considered issues of data coding and basic data description. These were important prerequisites to the topic of analyzing data. This chapter focuses on issues of analysis and, in particular, provides background information on statistical procedures commonly used in second language research.¹ We recommend that before conducting statistical analyses of data, researchers gain greater knowledge of statistics through academic coursework, statistical texts, or consultations with statistical experts.

9.2. DESCRIPTIVE STATISTICS

The first issue we deal with has to do with description and data display. Descriptive statistics can help to provide a simple summary or overview of the

¹This chapter deals with statistics rather than parameters. When researchers provide basic information about all members of a population (e.g., all first-year Spanish students at U.S. universities), they have information about the parameters of that population. It should be obvious that these data would be quite difficult to obtain; thus, researchers draw information from a representative subset (see chaps. 4 & 5 for a discussion of participant selection) of that population, known as a sample. The numerical information that we have about that population is referred to as statistics. In other words, our concerns are with the population, but our data come from samples.

data, thus allowing researchers to gain a better overall understanding of the data set. As Woods, Fletcher, and Hughes (1986) stated, "When a linguistic study is carried out, the investigator will be faced with the prospect of understanding, and then explaining to others, the meaning of the data which have been collected. An essential first step in this process is to look for ways of summarizing the results which bring out their most obvious features" (p. 8). In other words, because raw data are not particularly revealing, they must be organized and described in order to be informative. In this section we present an overview of three different types of descriptive statistics: measures of frequency, measures of central tendency, and measures of variability or dispersion. We also discuss ways of displaying this data visually to facilitate the exposition of summaries of findings.

9.2.1. Measures of Frequency

Measures of frequency are used to indicate how often a particular behavior or phenomenon occurs. For example, in second language studies, researchers might be interested in tallying how often learners make errors in forming the past tense, or how often they engage in a particular classroom behavior. One of the most common ways to present frequencies is in table format. For example, in Table 9.1, we present a sample frequency table from Storch and Tapper (1996), who provided the frequencies of different types of annotations that second language writers made on their own texts, indicating the areas in which they felt they were having difficulty.

In addition to tables, frequencies may also be represented graphically in forms such as histograms, bar graphs, or frequency polygons. In these graphic representations, the categories are typically plotted along the horizontal axis (*x*-axis), whereas the frequencies are plotted along the vertical axis (*y*-axis). For example, if we were to convert Storch and Tapper's (1996) frequency table into a graphic representation, one possible way would be through the bar graph seen in Fig. 9.1.a or with the same data through a line graph in Fig. 9.1.b.

Frequencies, as well as measures of central tendency (described later in this chapter), are often presented in second language studies even when they do not relate directly to the research questions. This is because frequency measures provide a succinct summary of the basic characteristics of the data, allowing readers to understand the nature of the data with minimum space expenditure. Also, frequencies and measures of central tendency can help researchers determine which sorts of statistical analyses are appropriate for the data.

TABLE 9.1
Sample Frequency Table

<i>Content of Student Annotations</i>	<i>Number</i>	<i>Total</i>
<i>Syntactic</i>		
Preposition or verb + preposition	21	
Verb tense	17	
Word order/sentence structure	17	
Articles	10	
Singular/plural agreement	8	
Word form	6	
Other	37	
		116
<i>Lexical</i>	70	70
<i>Blanket requests</i>		
Tenses	13	
Grammar	7	
Sentence structure	6	
Punctuation	4	
Other	29	
		59
<i>Discourse organization</i>	5	5
<i>Ideas</i>	5	5
Total		255

Source: Storch, N., & Tapper, J. (1996). Patterns of NNS student annotations in identifying areas of concern in their writing. *System*, 24(3), 329. Copyright © 1996 by Elsevier Science Ltd. Reprinted with the permission of Elsevier Science Ltd.

In order to visualize trends in the data, it is generally useful to plot the data even before carrying out statistical analysis. In this section, we have shown various ways of visually representing data (e.g., line graphs, bar graphs); these and other visual means of representation are useful in order to provide an impression of the data. For example, creating a scatterplot to assist with visualization of a dataset (see correlation figures in sec. 9.12.1)

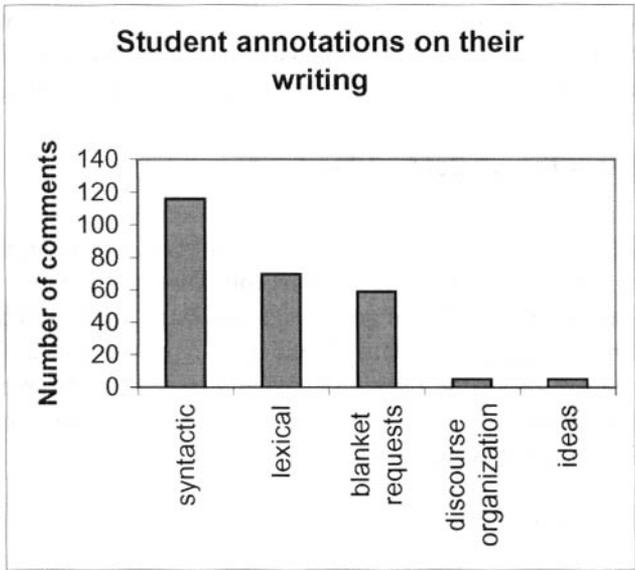


FIG. 9.1.a. Sample frequency bar graph.

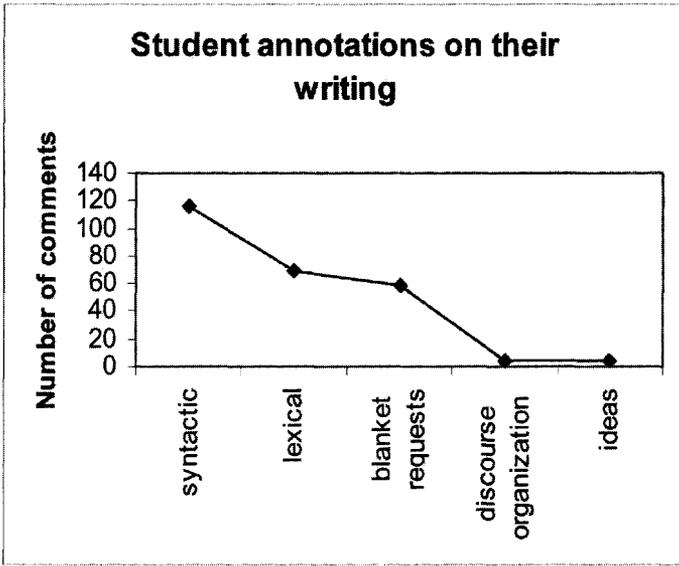


FIG. 9.1.b. Sample line graph.

can provide an early picture of any outliers in the data. Providing visual representations of results in graphical form can also contribute to a clearer understanding of any patterns confirmed through statistical testing.

9.2.2. Measures of Central Tendency

Although simple frequencies are useful ways of providing an initial picture of the data, they are not as precise as other measures, particularly when the data are obtained from different groups. Second language researchers often use one or more measures of central tendency to provide precise quantitative information about the typical behavior of learners with respect to a particular phenomenon. There are three commonly used measures of central tendency, each of which is discussed next.

9.2.2.1. Mode

Arguably the easiest measure of central tendency to identify is the mode. Simply put, the mode is the most frequent score obtained by a particular group of learners. For example, if the ESL proficiency test scores recorded for a group of students were 78, 92, 92, 74, 89, and 80, the mode would be 92 because two students in this sample obtained that score. Although this measure is convenient in that it requires no calculations, it is easily affected by chance scores, especially if the study has a small number of participants. For this reason, the mode does not always give an accurate picture of the typical behavior of the group and is not commonly employed in second language research.

9.2.2.2. Median

Another measure of central tendency that is easy to determine is the median. The median is the score at the center of the distribution—that is, the score that splits the group in half. For example, in our series of ESL proficiency test scores (78, 92, 92, 74, 89, 80), we would find the median by first ordering the scores (74, 78, 80, 89, 92, 92) and then finding the score at the center. Because we have an even number of scores in this case (namely, six), we would take the midpoint between the two middle scores (80 & 89), or 84.5. This measure of central tendency is commonly used with a small number of scores or when the data contain extreme scores, known as outliers (see sec. 9.2.2.4 for an explanation of outliers).

9.2.2.3. Mean

The most common measure of central tendency is the mean, or the arithmetic average.² Furthermore, because the mean is the basis for many advanced measures (and statistics) based on group behavior, it is commonly reported in second language studies. For our scores (78, 92, 92, 74, 89, 80), the mean would be the sum of all scores divided by the number of observations, (or $\Sigma x/n =$) 84.2. It should be kept in mind that even though the mean is commonly used, it is sensitive to extreme scores, especially if the number of participants is small.

The mean may be represented visually through the use of graphics, including the bar graph. For example, Toth (2000) created the graph in Fig. 9.2 for his study of the role of instruction, L2 input, and universal grammar in the acquisition of the Spanish morpheme *se* by English-speaking adult learners. In this graph, he provided a visual representation of the means of three different groups on three acceptability judgment tests (pretest, posttest, and delayed posttest). This visual presentation using a bar graph succinctly summarizes the information.

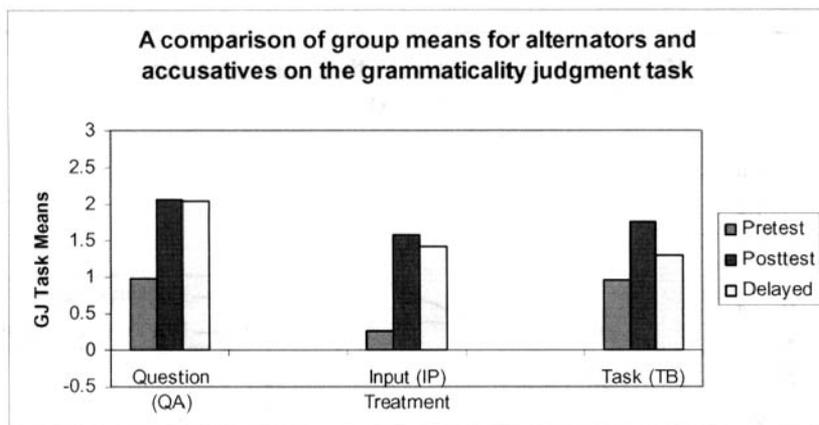


FIG. 9.2. Visual presentation of group means: Bar graph. Source: Toth, P.D. (2000). The interaction of instruction and learner-internal factors in the acquisition of L2 morphosyntax. *Studies in Second Language Acquisition*, 22(2), 189. Copyright © 2000 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

²Butler (1985) suggested, "The 'mean' is what the layman means by an average although the statistician would regard all three measures of central tendency as types of average" (p. 27).

Alternatively, means may be shown through the use of a line graph. For example, Zsiga (2003) compared patterns of consonant-to consonant timing at word boundaries in Russian and English to investigate the roles of transfer and the emergence of linguistic universals in second language articulation. Zsiga provided the graph in Fig. 9.3 to illustrate a significant interaction between L1 and language spoken, showing that the articulatory timing patterns of native Russian and English were different. The graph shows the mean duration ratios for English speakers and Russian speakers speaking their L1s and L2s, respectively.

In terms of measures and displays of central tendency and summaries of the data, it is always important to be flexible to the needs of your particular research questions and data set. In the words of Woods et al. (1986):

Although it will usually be possible to display data using one of the basic procedures ..., you should always remain alive to the possibility that rather special situations may arise where you may need to modify or extend one of those methods. You may feel that unusual data require a rather special form of presentation. Remember always that the major purpose of the table or graph is to communicate the data more easily without distorting its general import. (pp. 20–21).

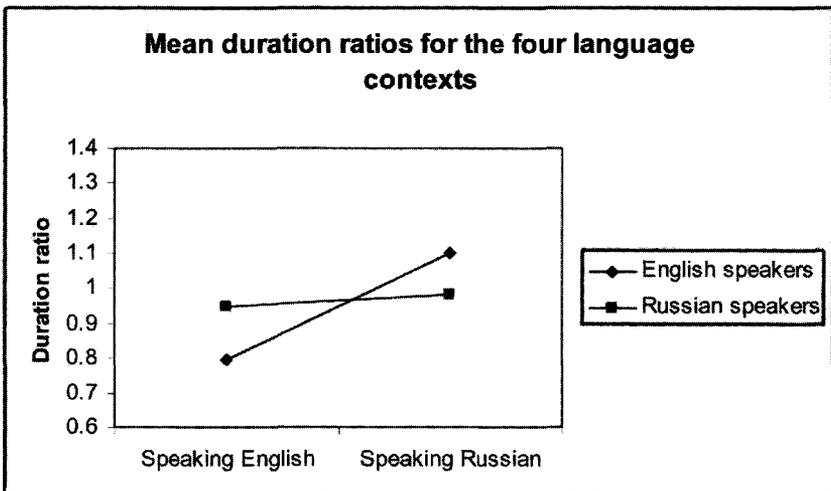


FIG. 9.3. Visual presentation of means: Line graph. Zsiga, L. (2003). Articulatory timing in a second language. *Studies in Second Language Acquisition*, 25(3) 413. Copyright © 2003 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

9.2.2.4. Outliers

Earlier in this section we mentioned the concept of outliers. These represent data that seem to be atypical of the rest of the dataset. The presence of outliers strongly suggests that the researcher needs to take a careful look at the data and determine whether the data collected from specific individuals are representative of the data elicited from the group as a whole. There are times when researchers may decide not to include outlier data in the final analysis, but if this is the case there needs to be a principled reason for not including them beyond the fact that they “don’t fit right.” Should researchers decide that there are principled reasons for eliminating outlying data, a detailed explanation in the research report needs to be provided. Following are two hypothetical examples in which a researcher might, after careful consideration, decide to eliminate some data.

Example 1:

- Data elicitation:** Sentence matching (see chap. 3 for further discussion of this elicitation technique). Participants are instructed to press the Yes button (the J key on the keyboard) if the sentences match or the No button (the F key on the keyboard) if the sentences do not match.
- Problem:** One participant has: (a) pressed only the Yes button throughout the experiment for all the sentences, and (b) consistently pressed it very quickly (i.e., the reaction times are much faster than the average).
- Possible reason:** Participant was not attentive to the task and repeatedly pressed only one button, suggesting that there was little processing going on.
- Decision:** Delete this individual’s data.
- Justification:** These data did not represent the processing that one has to assume for sentence matching.

Example 2:

- Data elicitation:** Child–child interactions in which the researcher is measuring feedback provided by children to their peers.
- Problem:** One child’s behavior appears to be unlike the others in that no feedback is ever provided.
- Further exploration:** In talking to the teacher, it was found that the child had a severe learning disability. It was also typical for this child not to stay on task in other classroom activities.

Decision:	Delete this individual's data.
Justification:	This child was most likely not on task. This child did not represent the population from which data were being collected.

Both of these examples are based on data that appeared to be unlike the rest of the dataset. The data were not immediately deleted, but because they were outliers, the researchers took a closer look at what might have been going on. It was only after a careful consideration and a determination that these data did not reflect a valid characterization of the construct of interest that the researchers decided that it was appropriate not to include the data in the final data pool. As mentioned earlier, information and justification of decisions like this should be included in a final research report.

Examples 1 and 2 illustrate occasions when it may be necessary to eliminate all of an individual's data when the extent to which they were on task is questionable. There are also cases when it may be appropriate to remove a subset of the data. For example, Duffield and White (1999) excluded from analysis those responses on a sentence-matching task from any participant who had an overall error rate of greater than 15% (e.g., said sentences were different when they were actually the same, and vice versa). Or, in some reaction time experiments, such as the one described in Example 1, a researcher might eliminate responses greater than a certain length of time, known as a cutoff point (e.g., 5000msec; Lotto & de Groot, 1998). A researcher might also move responses longer than the cutoff time to that cutoff point. For example, Duffield and White (1999) calculated the mean response time on a sentence-matching task for each individual. All responses that "fell outside a cut-off of ± 2 standard deviations of a particular subject's personal mean were corrected to the corresponding cut-off value" (p. 145). In other words, there was a maximum response time value that was used in their analysis.

9.2.3. Measures of Dispersion

Measures of central tendency are useful in that they give both the researcher and the reader an idea of the typical behavior of the group. However, the use of measures of central tendency alone may also obscure some important information. For instance, consider the hypothetical case of two groups of learners who take a final exam. One group of students obtains scores of 45, 99, 57, 17, 63, and 100, whereas the other group obtains scores

of 66, 62, 65, 64, 63, and 60. Both groups have approximately the same mean (63.5 & 63.3, respectively). However, if you report only the mean, you will not be able to show that the groups have a fairly different dispersion of scores: One group's scores are all close to the mean; the other group's scores are more widely dispersed. How can we present this additional information on the dispersion, or variability, of scores?

One informal way to do so is by presenting the range of scores. The range is the number of points between the highest and lowest scores on the measure. For example, the range for the first group of test scores would be 83 (17–100), whereas the range for the second would be 6 (60–66). The range, although easy to calculate, is not commonly reported in second language studies because it is sensitive to extreme scores and thus is not always a reliable index of variability.

A more common way of measuring variability is through the calculation of the standard deviation. Simply put, the standard deviation is a number that shows how scores are spread around the mean; specifically, it is the square root of the average squared distance of the scores from the mean. In other words, one takes the differences between each score and the mean and squares that difference. The next step is to add up these squared values, and divide by the sample size. The resulting number is called the variance. The standard deviation is the square root of the variance. As an example, consider the scores given earlier: 45, 99, 57, 17, 63, 100. To calculate the standard deviation, the following steps are taken:

1. Calculate the mean.

$$\Sigma x/n = 63.5$$

2. Subtract the mean from each score and square the difference. $(x - \bar{x})^2$

<i>Score</i>	<i>Mean</i>	<i>Difference</i>	<i>Difference Squared</i>
49	63.5	-14.5	210.25
99	63.5	35.5	1260.25
57	63.5	-6.5	42.25
17	63.5	-46.5	2162.25
63	63.5	-0.5	0.25
100	63.5	36.5	1332.25

3. Sum the differences squared and divide by the number of scores (6) to arrive at variance.

Score	Mean	Difference	Difference Squared
49	63.5	-14.5	210.25
99	63.5	35.5	1260.25
57	63.5	-6.5	42.25
17	63.5	-46.5	2162.25
63	63.5	-0.5	0.25
100	63.5	36.5	1332.25
			$\Sigma=5007.5$

$$\text{Variance} = 834.58$$

4. Take the square root of the variance.

$$\text{SD} = 28.89$$

The second set of scores given earlier (66, 62, 65, 64, 63, 60) are closer to one another. If we do the same calculation as previously, we see that the variance is 3.89 and the standard deviation is 1.97. Thus, although the means are similar, the amount of dispersion from the mean is quite different.

The larger the standard deviation, the more variability there is in a particular group of scores. Conversely, a smaller standard deviation indicates that the group is more homogeneous in terms of a particular behavior. We return to standard deviations later in our discussion of normal distributions.

Because the mean does not provide information about how scores are dispersed around the mean, the standard deviation (SD) should always be reported in second language research, often in a table along with the mean (M) and the number of subjects (n). An example of a table (Table 9.2) with this information comes from Rodríguez and Abreu (2003), who investigated the construct of *anxiety* in preservice teachers (native speakers of Spanish) majoring in English and French. The teachers were at different proficiency levels and came from two universities. Table 9.2 presents only the results from one of the universities in their example. The table portrays descriptive information.

One can examine the SDs and means in relation to one another. All groups (with the exception of Level 5 for French Anxiety) are more or

TABLE 9.2
Sample Mean and Standard Deviation Table:
English and French Anxiety Score by ... Level, Restricted Sample

Level	English Anxiety			French Anxiety		
	M	SD	n	M	SD	n
1	74.42	14.87	12	76.75	17.32	12
3	90.42	14.98	12	89.08	13.99	13
5	94.38	14.00	8	93.50	120.52	8

Note: Maximum score = 165.

Source: Rodríguez, M., & Abreu, O. (2003). The stability of general foreign language classroom anxiety across English and French. *The Modern Language Journal*, 87, 371. Copyright © 2003 by Blackwell. Reproduced with the permission of Blackwell.

less equally dispersed from the means. If SDs are consistently large compared to the mean, you have groups with little homogeneity. In general, researchers should closely examine data with SDs that are consistently larger than the mean. Measures of dispersion (particularly standard deviations) can serve as a quality control for measures of central tendency; the smaller the standard deviation, the better the mean captures the behavior of the sample.

As was the case with frequencies, this information can also be represented visually. For example, Robinson (1997) provided the graph shown in Fig. 9.4 in his study of the effect of different instruction conditions on the ability of adult Japanese ESL learners to acquire a rule about verbs. In this graph, the mean scores are represented by the height of the bars, whereas the black line extending from on top of each bar represents the size of the standard deviation.

As we noted earlier, it is important to include measures of variability in descriptions of data. Each offers different information, but when taken together they provide a richer understanding of the data than when viewed alone. As seen later in this chapter, means and standard deviations figure prominently in many statistical analyses.

9.3. NORMAL DISTRIBUTION

A distribution describes the clusterings of scores/behaviors. In a normal distribution (also known as a bell curve) the numbers (e.g., scores on a par-

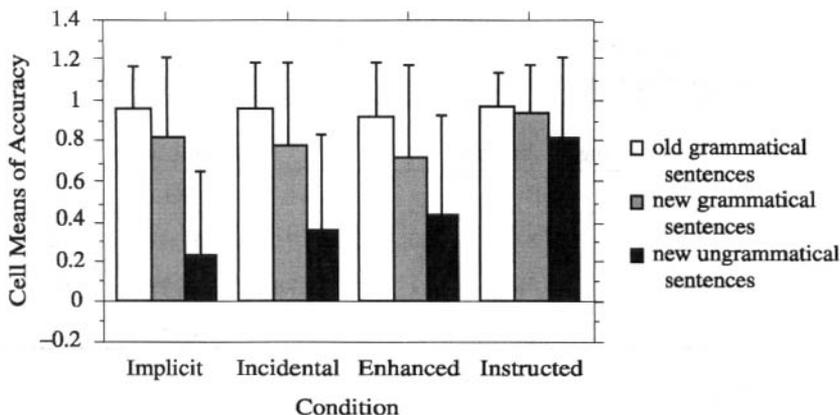
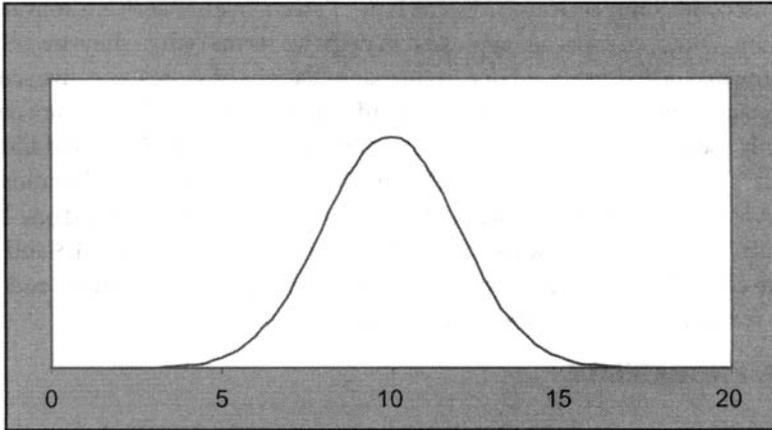


FIG. 9.4. Sample visual representation of mean and standard deviation. Source: Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced and instructed conditions. *Studies in Second Language Acquisition*, 19, 235. Copyright © 1997 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

ticular test) cluster around the midpoint. There is an even and decreasing distribution of scores in both directions. Figure 9.5 shows a normal distribution. As can be seen, the three measures of central tendency (mean, mode, median) coincide at the midpoint. Thus, 50% of the scores fall above the mean and 50% fall below the mean. Another characteristic of a normal distribution relates to the standard deviation. In a normal distribution, approximately 34% of the data lie within 1 standard deviation of the mean. In other words, 34% of the data are one standard deviation above the mean and 34% are one standard deviation below the mean. We have now described 68% of the data. If we look at two standard deviations above and below the mean, we capture an additional 27% for a total of 95%. Thus, only 5% of the data in a normal distribution lies beyond 2 standard deviations from the mean. Finally, approximately 2.13% of the data fall between 2 and 3 standard deviations, leaving only approximately .3% of the data beyond 3 standard deviations above and below the mean. If we know that a group of scores is normally distributed and if we know the mean and the standard deviation, we can then determine where individuals fall within a group of scores. Many statistics assume normal distribution of scores. Figure 9.5 represents a normal distribution with the mean, mode, and median corresponding at the midpoint.



Mean= 10

Mode=10

Median=10

FIG. 9.5. Normal distribution.

9.4. STANDARD SCORES

There are times when we want to compare an individual's performance on different tests. For example, we might want to compare a score on a vocabulary test with a score on a test of grammar. Given the nature of the two tests, it is also likely that the maximum score on each is different. It would, of course, not be prudent to compare a score of 22 on one test with a score of 22 on another when one was based on a total possible score of 75 and the other based on a total possible score of 25. One way to make a more meaningful comparison is to convert these raw scores into standard scores.

The two most common standard scores are z scores and T scores. The first type, z scores, uses standard deviations to reflect the distance of a score from a mean. If a score is one standard deviation above the mean, it has a z score of +1, a score that is two standard deviations above the mean has a z score of +2, and a score that is one standard deviation below the mean has a z score of -1. The calculation of a z score is straightforward: We subtract the mean from the raw score and divide the result by the standard deviation. The formula is given in Appendix J.

A second common standard score is the *T* score. In essence, it is a converted *z* score. Often, *z* scores are expressed in negative terms (when they are below the mean) and in fractions. For certain manipulations of scores, negative scores are inappropriate. If nonnegative standard scores are needed, *T* scores are commonly used. *T* scores are calculated by multiplying the *z* score by 10 and adding 50 ($(z \times 10) + 50$). Consider a test with a mean of 60 and a standard deviation of 14. A learner who receives a score of 39 on this test has scored one and one-half standard deviations below the mean, and would have a *z* score of -1.5 and a *T* score of 35. The relationship between means and standard deviations and the two standard scores discussed here can be seen in Fig. 9.6.

9.5. PROBABILITY

The purpose of conducting statistical tests is to provide information about the likelihood of an event occurring by chance. The probability value (referred to in research reports as the *p*-value) that is reported is designed to provide confidence in the claims that are being made about the analysis of the data. We are all familiar with the concept of probability from everyday life. Insurance companies rely on the concept of probability to determine rates using various factors such as age, and health (for life insurance) or age and driving record (for automobile insurance). The general way of expressing probability is through a percentage (.20=something occurring 20% of

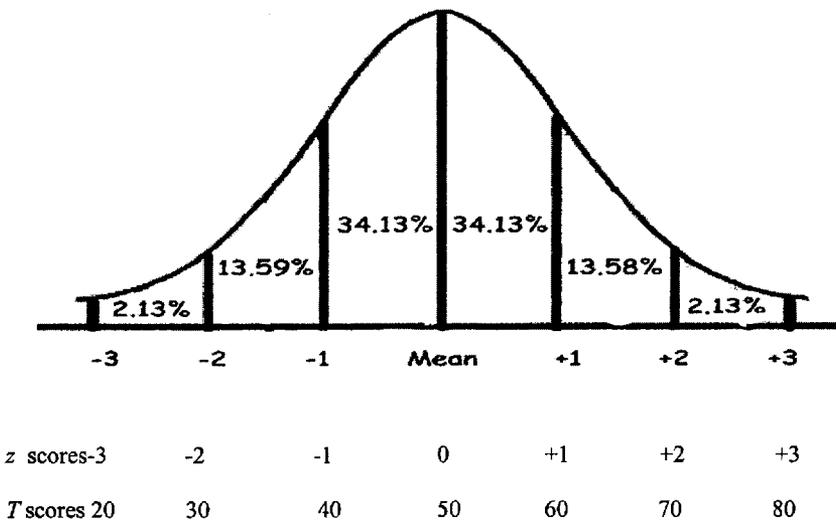


FIG. 9.6. Means, standard deviations, *z* scores, and *T* scores.

the time). Probability is an expression of the likelihood of something happening again and again. For example, if the probability is .05, there is a 5% possibility that the results were obtained by chance alone. If the probability is .50, there is a 50/50 possibility that the results were obtained by chance. The accepted *p*-value for research in second language studies (and in other social sciences) is .05. A *p*-value of .05 indicates that there is only a 5% probability that the research findings are due to chance, rather than to an actual relationship between or among variables. In second language research reports, probability levels are sometimes expressed as actual levels and sometimes as simply greater or less than .05 or some other probability level. Table 9.3 shows actual *p*-values from a study on planning and focus on form with L1 English speakers learning Spanish (modified from Ortega, 1999). The column labeled *F*-value reflects the specific statistical procedure that Ortega used, analysis of variance (ANOVA), and is discussed later in this chapter. Table 9.4, from a study on word meaning by L1 English speakers learning Spanish (modified from Barcroft, 2003), shows *p*-values being expressed in relation to .05 and .01.

In chapter 4 we introduced the concept of null hypotheses. Null hypotheses predict that there is no relationship between two variables; thus, the statistical goal is to test the hypothesis and reject the null relationship by showing that there is a relationship. Let us take the following hypothesis: "Resumptive pronouns (*The man that I saw him is very smart*) will decrease with time." This

TABLE 9.3.
Example of Expression of Probability Levels:
Summary of Findings From ANOVAs on IL Measures

<i>Measure</i>	<i>F-Value</i>	<i>p-Value</i>
Words per utterance	8.444	.0002
Noun-modified TLU	5.8472	.0217
Pruned speech rate	16.0625	.0004
Type-token ratio	1.5524	.2221
Article TLU	4.3599	.0451

Source: Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 126. Copyright © 1999 by Cambridge University Press. Reprinted with the permission of Cambridge University Press.

TABLE 9.4
Example of Expression of Probability Levels: Repeated Measures ANOVA for Effect of Condition and Time on Cued Recall

<i>Source</i>	<i>F</i>
Time	4.84*
Condition	9.06**

* $p < .05$, ** $p < .01$.

Source: Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *The Modern Language Journal*, 87, 557. Copyright © 2003 by Blackwell. Reproduced with the permission of Blackwell.

hypothesis predicts change in a particular direction, that is, the occurrence will decrease over time. We could express this hypothesis as a null hypothesis as follows: "There is no relationship between the use of resumptive pronouns and the passage of time." We can then test whether the null hypothesis can be rejected. Consider the following hypothetical scenarios representing the number of instances of null subject use over time:

	<i>Time 1</i>	<i>Time 2</i>	<i>Time 3</i>	<i>Time 4</i>
Scenario 1	4	2	2	1
Scenario 2	30	20	8	1

We can see that the difference in the number of instances in Scenario 1 is slight, suggesting that this may be a random finding and that were we to repeat this study many times, the results would be different. If we were to do a statistical test, we would probably come up with a high p -value and we would have little confidence that our results would be the same were the test to be repeated. On the other hand, the difference in the numbers in Scenario 2 is such that we would have more confidence in our results not being due to chance alone. A low level of probability would indicate this.

As noted previously, probability is an estimation of the likelihood of something occurring due to chance. Two potential problems with such estimates, commonly referred to as Type I and Type II errors, are noteworthy. Type I errors occur when a null hypothesis is rejected when it should not

have been rejected; Type II errors occur when a null hypothesis is accepted when it should not have been accepted. Examples include the following:

<i>Error Type</i>	<i>Definition</i>	<i>Example</i>
Type I	Reject null hypothesis when it should not be rejected.	A statistical test shows a significant difference between an experimental and a control group ($p < .05$), and the researcher confirms that a treatment has been successful when in actuality it was unlikely that the two groups were different.
Type II	Accept null hypothesis when it should not be accepted.	A statistical test shows no significant difference between an experimental and a control group, and the researcher confirms the treatment has not been successful when there really was a difference.

Before moving to a discussion of statistics, we utter a word of caution about the difference between significance and meaningfulness. When we have a large sample size, it is often not difficult to get statistical significance. Assume, for example, that you are testing the effect of recasts versus models for the learning of irregular past tense verbs in English. Assume also that you have a sample size of 500 learners in the recast group and 500 in the model group (a highly unlikely event in second language research). Following the treatment, you give a posttest, and the model group has a mean score of 8.8 and the recast group has a score of 9.1. With such a large sample size it is possible that this difference is significant, but given the small difference (.3) we might not want to make strong claims based on these results. However, in second language research we generally deal with much smaller sample sizes, making it difficult to get statistical significance. The commonly accepted level for significance in second language research is .05. This is known as the alpha (α) level; different alpha levels can be set by the researcher at the onset of the research. For certain research—for example, when high-stakes decisions will not be based on the analysis—the researcher may decide to set a less conservative alpha level. In fields such as medicine where the stakes are high, the alpha levels are much more conservative, thereby reducing the likelihood of chance occurrences. In sum, the p -value is the exact probability level matching the calculated statistic. The actual p -value must be lower than the predetermined alpha level for the results of the analysis to be considered significant. In second language research, even when the alpha level of .05 is

used, researchers occasionally describe their findings in terms such as “approaching significance” or “demonstrating trends” when the p -value is between .05 and .075 or even .10.

In considering the difference between meaningfulness and significance (in the statistical sense), we need to recognize that second language learning is a slow and complex process often involving a period of production of correct forms only to be followed by a later period of production of incorrect forms. Therefore, we often need longer periods of observation, but the exigencies of research do not often allow long periods of time. Thus, it may be that meaningful trends are worthy of discussion, independent of statistical significance. As Gass et al. (1999) noted:

The need to have all results adhere to a .05 standard may be questionable. Shavelson (1988) noted that the convention of using .05 or .01 “grew out of experimental settings in which the error of rejecting a true H_0 was very serious. For example, in medical research, the null hypothesis might be that a particular drug produces undesirable effects. Deciding that the medicine is safe (i.e., rejecting H_0) can have serious consequences. Hence, conservatism is desired” (p. 248).

Shavelson went on to say that “often in behavioral research...the consequences are not so dire” (p. 248). He suggested that a level of .25 might be appropriate in some cases. He concluded his discussion by saying that “some wisdom, then, should be exercised in setting the level of significance” (p. 248) and pointed out that there is a trade-off between the level of significance that one sets and the power of one’s conclusions.

Given the essential arbitrariness in setting significance levels and given the constraints in conducting [second language research, particularly] classroom research, we feel that trends are important and at least point to the notion that experiments should be replicated, particularly when it is impractical or impossible for experiments to cover a long period. We also believe that trends may at times be as meaningful as statistical significance. (pp. 575–576)

We are not suggesting that different levels or standards for significance should apply to second language research than those that apply to education, social or cognitive sciences in general; what we are suggesting is that given the nature of second language research, it is not always necessary to completely discount trends in all data that do not fit within the narrow confines of the standard alpha level of .05.

9.6. INFERENCE STATISTICS

The goal of some types of second language research is to go beyond uncovering information about how a particular group of students—for example, those enrolled in first-year Spanish—learn a particular part of the language. Rather, the goal is to generalize beyond the results. In other words, such researchers want to make inferences from the particular sample to the population at large. Given that it is impossible to gather data from all members of the population, inferential statistics can allow researchers to generalize findings to other, similar language learners; that is, to make inferences. In the following sections we deal with some of the most common inferential statistics that are used in applied linguistics and second language research.

9.6.1. Prerequisites

Before moving to present information about specific statistical analyses, we briefly discuss some basic concepts that relate to statistical procedures. Although the first two—standard error of the mean and standard error of the difference between sample means—are not concepts that are presented in research reports, they are important for conceptualizing the statistics presented later in the chapter.

9.6.1.1. *Standard Error of the Mean*

Standard error of the mean (SEM) is the standard deviation of sample means. The SEM gives us an idea of how close our sample mean is to other samples from the same population. If we know that the mean for the total population is 50 and if we know that the SEM is 5, we also know that if our sample mean is 52, it is within one SEM of the population mean and is within 34% of all sample means taken from the population. The formula for the calculation of SEMs is presented in Appendix J. Because we do not know the mean for the total population, this is not a precise measure, but it is important in determining the standard error of the difference between sample means, discussed in the next section.

9.6.1.2. *Standard Error of the Difference Between Sample Means*

Standard error of the difference between sample means (SED) is based on the assumption that the distribution of differences between sample

means is normal. This distribution, because it is normal, will have its own mean and standard deviation. This standard deviation is known as the SED. In order to calculate the SED, one needs to know the SEM of the two samples in question (app. J).

9.6.1.3. Degrees of Freedom

The concept of degrees of freedom is necessary as we consider the determination of significance of statistical tests. To put it simply, the degree of freedom is the number of scores that are not fixed. Suppose we know that our total value on a test adds up to 50 and we have 5 scores contributing to this value of 50. If we know what 4 of the scores are, the 5th one is fixed; it cannot vary. In other words, only one of the scores cannot vary. In this case, 4 represents the degrees of freedom. This is important when we look up critical values on statistical tables. Statistics tables are organized by alpha level (e.g., $p < .05$) and degrees of freedom and are expressed in terms of critical values.

9.6.1.4. Critical Values

This is the value that we can use as a confidence measure to determine whether our hypothesis can be substantiated. When a statistic is calculated, the numerical result of the calculation is compared against the statistical table, to determine whether it reaches the critical value. If the result of the calculation reaches or surpasses the appropriate value, the findings are considered statistically significant. Researchers can look up critical values in a statistics table, although statistical packages that calculate statistics also provide the critical value. This is further discussed later in this chapter.

9.6.1.5. One-Tailed Versus Two-Tailed Hypotheses

When we discussed hypotheses in chapter 1, we presented some hypotheses that predicted differences in one direction or another and others that were neutral as to direction, that is, they predicted a difference but not in which direction the difference was expected. The former (those that predict a difference in one direction) are known as one-tailed hypotheses and require a different critical value than do the “neutral” or two-tailed hypotheses. Examples of one-tailed and two-tailed hypotheses follow:

One-tailed hypothesis: The group that received explicit grammar instruction before reading a passage with those grammatical elements will have higher comprehension scores than will those who had vocabulary instruction before reading a passage with those vocabulary items.

This hypothesis clearly predicts which group will perform better.

Two-tailed hypothesis: The group that received explicit grammar instruction before reading a passage with those grammatical elements will have a different level of comprehension than will those who had vocabulary instruction before reading a passage with those vocabulary items.

This hypothesis predicts a difference in the performance of the two groups, but says nothing about which group will perform better.

9.6.2. Parametric Versus Nonparametric Statistics

There are two broad categories of inferential statistics known as parametric and nonparametric tests. As the names suggest, they deal with the parameters of the population from which researchers have drawn samples.

With parametric statistics, there are sets of assumptions that must be met before the tests can be appropriately used. Some of the assumptions for parametric tests include the following:

- The data are normally distributed, and means and standard deviations are appropriate measures of central tendency.
- The data (dependent variable) are interval data (e.g., scores on a vocabulary test; see chap. 4 for further information).
- Independence of observations—scores on one measure do not influence scores on another measure (e.g., a score on an oral test at Time 1 does not bias the score on an oral test at Time 2).

Again, we refer the reader to detailed descriptions in statistics books related to the specific sets of assumptions for each test.

The assumptions underlying nonparametric tests are minimal. Nonparametric tests are generally used with frequency data (e.g., the amount of other-correction in class discussion in different classrooms)

or when the assumptions for parametric tests are not met. Parametric tests have more power. This means that they are more likely to detect a genuine effect because they are more sensitive. Parametric tests are also more likely to detect an effect that does not really exist. One reason for the greater power of parametric tests is that there is more information that feeds into the statistic. If a statistical test lacks power, it may be difficult to detect the effect of the independent variable upon the dependent variable, resulting in a Type II error, or failure to reject the null hypothesis when it is incorrect. However, using a parametric statistic when it is not appropriate can lead to a Type I error, an incorrect rejection of the null hypothesis.

In the following sections, we briefly discuss some of the more frequently used parametric and nonparametric tests used in second language research.

9.6.3. Parametric Statistics

In this section we deal with t-tests and analysis of variance.

9.6.3.1. t-tests

The t-test can be used when one wants to determine if the means of two groups are significantly different from one another. There are two types of t-tests—one is used when the groups are independent and the other, known as a paired t-test, is used when the groups are not independent, as in a pre-test/posttest situation when the focus is within a group (a person's performance before treatment compared with his or her own performance after treatment). Following are examples of types of research in which a t-test and a paired t-test would be appropriate:

Example 1:

Description: You have completed a research study looking at the effectiveness of two kinds of feedback on learners' vocabulary test scores. Group 1 has 35 learners and Group 2 has 33. You have calculated the means and standard deviations of the end of semester exams of the two groups and you believe that all of the assumptions for a parametric test have been met.

Statistic: You compare the two groups using a t-test.

Example 2:

- Description:** You have conducted a study on the effectiveness of a particular way of teaching reading. You have given a pretest and a posttest. Each individual has two scores (pretest and posttest). You want to know if the improvement following the treatment was significant.
- Statistic:** A paired t-test is appropriate; each person is paired with him- or herself on the two tests.

Example 3:

- Description:** You have conducted a study on the acquisition of relative clauses by Korean and Spanish learners of English. There are two groups, matched for native language and gender. Group 1 consists of 3 male native speakers of Korean, 4 female native speakers of Korean, 5 male native speakers of Spanish, and 4 female speakers of Spanish. Group 2 has the same profile. (Groups could be matched on a variety of factors, e.g., age, preexperiment tests, reading tests, listening tests, etc.) Group 1 receives instruction on subject relative clauses; Group 2 receives instruction on indirect object relative clauses. You have pretest scores and posttest scores for each individual on a range of relative clause types and calculate a gain score for each. You want to see if there are differences in learning between Groups 1 and 2.
- Statistic:** Paired (matched) t-test is appropriate because you have matched pairs. That is, Korean Male #1 in Group 1 can be compared with Korean Male #1 in Group 2, and so forth.

A word of caution about the use of t-tests is necessary. As noted, they are appropriate when comparing two groups, but there is a tendency in second language research to run t-tests on different parts of a data set in a way that is overusing the test. Using an alpha level of .05 means that there is a 5% possibility of getting significance by chance. In other words, 1 time out of 20 we might have a significant result that in actuality should not be significant. If one carries out 10 t-tests, for example, the odds are increased that a Type I

error will be produced. For example, if you have conducted an experiment in which there were four groups, carrying out multiple two-way comparisons or multiple t-tests on subparts of the data (e.g., native speakers of one language versus native speakers of another language, or males versus females) could be considered overusing the test. If there are multiple groups, rather than doing multiple two-way comparisons using t-tests, another statistic, such as an analysis of variance, may be more appropriate because analysis of variance calculations mathematically account for the increased chance of error that occurs as multiple comparisons are made.

9.6.3.2. Analysis of Variance (ANOVA)

In the previous section we discussed t-tests, which enable researchers to compare performance on two groups. Many research designs require comparisons with more than two groups and ANOVA may be appropriate in this context. ANOVA results provide an *F* value, which is a ratio of the amount of variation between the groups to the amount of variation within the groups.

Example: You have conducted a study in which you are comparing the effectiveness of three different types of instruction. You are confident that the assumptions of a parametric test have been met. You want to compare the results and use an ANOVA to see if group differences are due to chance or are sufficient to reject the null hypothesis.

A sample result from an analysis of variance is presented in Table 9.5. This table includes the information that is relevant to understand an analy-

TABLE 9.5
Example of an ANOVA Results Table

Source of Variance	SS	df	MS	F
Between groups	521.43	2	260.71	53.98*
Within groups	202.66	42	4.83	
Total	724.09	44		

* $p < .01$

sis of variance result (see app. J for the formula). An ANOVA provides information on whether or not the three (or more) groups differ, but it provides no information as to the location or the source of the difference. That is, is Group 1 significantly different from Groups 2 or 3 or is Group 2 significantly different from Group 3? To determine the location of the difference when the F value is significant, a post-hoc analysis is used. Common post-hoc analyses include the Tukey test, the Scheffé test, and Duncan's multiple range test. A typical display showing the source of a difference for the possible study described in the previous example is presented in Table 9.6. In this hypothetical example, differences were found between the groups who had Instruction 1 and Instruction 2, and between the groups who had Instruction 1 and Instruction 3. No other differences were found; the differences between Instruction 2 and Instruction 3, for example, were not significantly different.

9.6.3.3. Two-way ANOVA

In second language research, there is often a need to consider two independent variables, for example, instruction type and proficiency level. When there is more than one independent variable, the results will show main effects (an effect of each independent variable without considering the effect of the other) and an interaction effect which is the effect of one independent variable that is dependent on the other independent variable. In Fig. 9.7 we can see that there is not a straightforward relationship between test scores and instruction type. Rather, there is an interaction between instruction type and proficiency level such that high proficiency students do

TABLE 9.6
Example of a Post-Hoc Table

Group	Group		
	Instruction 1	Instruction 2	Instruction 3
Instruction 1		*	*
Instruction 2	*		
Instruction 3	*		

* Pairs where there was a significant difference at the .05 level.

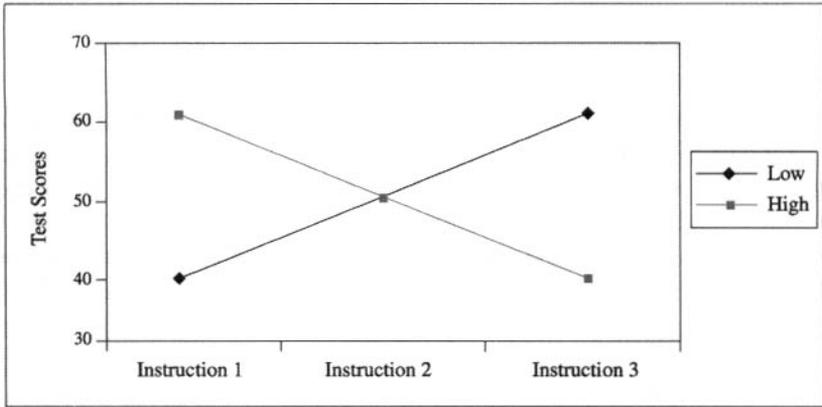


FIG. 9.7. Instruction type as a function of proficiency level.

better with instruction type 1, whereas low proficiency students perform better with instruction type 3.

9.6.3.4. Analysis of Covariance (ANCOVA).

There are times when there might be a preexisting difference among groups and the variable where that difference is manifested is related to the dependent variable. In other words, differences in means on variable X will show up on a pretest. The preexisting difference will need to be controlled for and is referred to as the covariate. Because of differences among groups, the posttest results will need to be adjusted. The amount of adjustment will depend on two factors: how large the difference is between the pretest means and the change between the pretest and the posttest (the dependent variable).

Example: You are testing three types of pedagogical treatments for learning the orthographic writing system of Arabic (explanation, visual repetition, practice). To do this you use three separate first-semester University-level Arabic classes. You have a pretest, treatment, posttest design. You find that your groups are not matched at the outset on language proficiency. Thus, your pretest score, the covariate, will have to be adjusted to compensate for the fact that one group starts at a higher level than the other. If no adjustment is made, we would not know whether the group with the initial higher score learned more or appeared to learn more because of the higher initial score. An ANCOVA is appropriate.

9.6.3.5. Multivariate Analysis of Variance (MANOVA)

The MANOVA is part of the family of analyses of variance. It differs from an ANOVA in that it has more than one dependent variable. In order to appropriately use a multivariate analysis of variance, there has to be justification for believing that the dependent variables are related to one another.

Example: You have conducted a study of the effectiveness of different interlocutor types on learner performance (as measured by oral abilities and grammar). You devise a spoken proficiency test as well as an acceptability judgment task to measure learning. Because you are interested in the relationship between oral abilities and grammatical knowledge for the different interlocutor types, a multivariate analysis of variance is appropriate.

9.6.3.6. Repeated Measures ANOVA

There are times when we might want to compare participants' performance on more than one task.

Example: You have conducted a study of different writing prompts on learner performance as measured by writing accuracy. You have developed a measure of accuracy that includes length of essay, error-free T-units, and sophistication of vocabulary. You have carefully devised a counterbalanced design in which each participant writes an essay under four conditions:

1. Timed with a prompt.
2. Untimed with a prompt.
3. Timed with no prompt.
4. Untimed with no prompt.

Because each individual does all the tasks, we have a repeated measures design. And, because we have three different sets of results to be compared, a repeated measures ANOVA is appropriate. A final point to consider is the determination of degrees of freedom. For t-tests, it is the number of participants in each group - 1. For analyses of variance, there is a between group and within group difference. The form is the number of groups (or independent samples) - 1; for the latter, degrees of freedom equals the total number of participants minus the number of groups.

9.6.4. Nonparametric Tests

As discussed earlier, nonparametric tests are generally used with frequency data or when the assumptions for parametric tests have not been met. In this section we discuss some of the most frequently used nonparametric tests in second language and applied linguistics research.

9.6.4.1. Chi Square (χ^2)

Chi square tests are often used with categorical (i.e., nominal) data. Examples of categorical data include age groups (e.g., 8–10 years old, 11–13, 14–16), gender, native language, types of relative clauses, and so forth. The chi square statistic relies on observed frequencies and expected frequencies.

Example: You want to determine whether ESL learners from different L1 backgrounds and of different genders are more likely to use stranded prepositions (*That's the man I talked to you about*). You elicit this structure from 40 learners with the same gender distribution (20 L1 Japanese and 20 L1 Spanish—10 males and 10 females each). You construct a table that looks like the following:

Participants Who Use Stranded Prepositions

	<i>Japanese</i>	<i>Spanish</i>
Male	8	4
Female	6	10

If native language and gender did not matter in the use of stranded prepositions, we would expect the values in each square in the table to be equal. You can determine whether the actual values are different from the expected values by using a chi square analysis. If the actual values differ from the expected values, it can be assumed that at least one of the variables (native language or gender) influences the use of stranded prepositions. The expected frequency is determined by taking the sum total of the observations (in this case, 28) and dividing it by the number of cells (in this case, 4). Hence, the expected frequency for each cell is 7. These are the values that feed into a chi square formula. Degrees of freedom are then determined by subtracting one from the number of columns. In this example, there is one degree of freedom. Degrees of freedom and corrections are generally built into computer programs that automatically calculate chi squares. When there is one degree of freedom, Yates' correction factor is often used.

Just as with many parametric statistics, chi square analyses rely on assumptions as to the type of data needed. Primary among these assumptions are the following:

- Each observation is independent; that is, it only falls in one cell. In the previous example, an individual is either male or female and is either Japanese or Spanish.
- The data are raw frequencies (not relative frequencies or percentages).
- Each cell has an expected frequency of at least 5.

The Fisher's exact test, a variant of the chi square test, may be more appropriate than a chi square in some situations, including those in which there are several cells with expected frequencies that are less than 5, or where there are cells with a value of 0.

When chi square tests are calculated to determine the relationship among several variables, the results will indicate significant relationships. However, as with ANOVA tests, the location of the significance is not identified. There are procedures (e.g., Haberman's residuals) that can be used to locate the significant relationships.

9.6.4.2. Mann-Whitney U/Wilcoxon Rank Sums

Other nonparametric tests are used with ordinal or interval data rather than categorical data. Mann-Whitney U and Wilcoxon Rank Sums are two such tests; we discuss these together as they are essentially the same test. These are comparable to the t-test in that they compare two groups but are used when the results are rank scores (i.e., with ordinal scale dependent measures). Both sets of scores are pooled and the scores are compared in relation to the median.

Example: You want to determine the effects of interaction on the ability of the interlocutor to comprehend descriptions. You design a study with two groups, each made up of 10 dyads: In one group interaction is allowed and in the other interaction is not allowed. You subsequently observe where objects were placed on a board (dependent variable, measure of comprehension). The object placement scores are quantified and converted into rank scores.

A Mann-Whitney U is appropriate because the interval data assumption of a parametric test was not met. Had that assumption been met, a t-test could have been used. Rather than degrees of freedom, the relevant information is the size of the larger and smaller sample.

9.6.4.3. Kruskal-Wallis/Friedman

A Kruskal-Wallis is a nonparametric test comparable to an ANOVA, but used when parametric test assumptions are not met. It is employed when a researcher wants to compare three or more independent groups. In other words, a between-groups comparison is being made. A Friedman test is the nonparametric counterpart to a repeated measures ANOVA. That is, when you have nonindependent samples and need to compare within groups, a Friedman may be appropriate. Degrees of freedom equals the number of samples minus 1.

In the preceding sections, we dealt with some commonly used parametric and nonparametric statistics in second language research. In Table 9.7 we summarize some of the different types of second language data together with possible statistical techniques.

9.7. STATISTICAL TABLES

In closing this discussion on parametric and nonparametric tests, we present two statistical tables to illustrate how they can be read and used. Most statistical textbooks include full versions of tables that can be consulted to determine if your test results are significant. If statistics are carried out using a computer-based statistical package (see later discussion), the results will be provided for you and there will be little need to consult a statistical table such as the ones given in this section. Table 9.8 provides a partial display of the distribution of *t*.

This table and other statistical tables display the minimum value (i.e., critical value—see sec. 9.6.1.4) based on the desired probability level and de-

TABLE 9.7
Summary of Statistics

<i>Type of Comparison/Type of Test</i>	<i>Parametric</i>	<i>Nonparametric</i>
Two independent samples	t-test	Mann-Whitney
Two related samples	Paired t-test	Wilcoxon
More than two independent samples	ANOVA	Kruskal-Wallis
More than two related samples	Repeated measures ANOVA	Friedman

TABLE 9.8
Distribution of t

<i>p</i> (<i>df</i>)	.1	.05	.02	.01	.001
11	1.796	2.201	2.718	3.106	4.437
12	1.772	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073

Note: *p* refers to the probability level; *df* refers to the degrees of freedom.

degrees of freedom that one must have to claim significance. There are two points to note about this table. First, to determine significance, one looks at the left-hand column at the relevant degrees of freedom. Second, one has to determine whether one has a one-tailed or a two-tailed hypothesis. The figures given across the top (*p* levels) are for a two-tailed hypothesis. For a one-tailed hypothesis, one halves the probability level. For example, Column 2 (headed by .1) is .05 for a one-tailed hypothesis. Thus, if one has 14 degrees of freedom on a two-tailed test, and has a value of 2.98, one can claim that the significance is $< .01$.

The second table (Table 9.9) is from a nonparametric test—chi square. The method for reading this table is the same as that for reading the t-test ta-

TABLE 9.9
Distribution of Chi Square

<i>p</i> (<i>df</i>)	.1	.05	.02	.01	.001
11	17.275	19.675	22.618	24.725	31.264
12	18.549	21.026	24.054	26.217	32.909
13	19.812	22.362	25.472	27.688	34.528
14	21.064	23.685	26.873	29.141	36.123
15	22.307	24.996	28.259	30.578	37.697

Note: *p* refers to the probability level; *df* refers to the degrees of freedom.

ble. If one has 15 degrees of freedom and one has a chi square value of 33.021, one has a significance level of $< .01$.

As mentioned earlier, many second language researchers will not often have to read a statistical table because most computer programs provide exact probability levels (e.g., in the form $p = .023$, rather than in the form of $p < .05$).

9.8. STRENGTH OF ASSOCIATION

There are times when we might want to determine how much of the variation is actually due to the independent variable in question (e.g., the treatment, the learner's language background, the learning context, etc.). That is, if we find a difference—for instance, in performance between native speakers of Japanese learning English and native speakers of Arabic learning English on some measure—we don't know how much of the difference is due to the fact that their native languages are different or to something else (which we probably cannot specify). The following sections discuss some statistical procedures that can help us address these questions.

9.9. ETA² AND OMEGA²

The most common measurement that can be used after a t-test is eta² (expressed as η^2), which goes beyond the fact that there is a significant difference and gives us an indication of how much of the variability is due to our independent variable. Consider Example 1 of a t-test in the study of two different types of vocabulary instruction. Suppose that the t-test indicates that the learners from Group 1 score significantly better on their end of semester exam than do the learners from Group 2. You know that there is a difference between these groups, but you don't know how much of that difference can be explained by the independent variable (instruction type). You calculate eta² and determine that $\eta^2 = .46$. That means that 46% of the variability in their scores can be accounted for by the instruction type.

The same reasoning applies for ANOVAs. Omega² (ω^2) is the statistic used when all groups have an equal n size. Otherwise eta² is appropriate. The formulæ for these tests are given in appendix J.

9.10. EFFECT SIZE

Effect size is a measure that gives an indication of the strength of one's findings. In 2000, the editor of the second language journal *Language Learning* (N.

Ellis) stated that “the reporting of effect sizes is essential to good research. It enables readers to evaluate the stability of research across samples, operationalizations, designs, and analyses. It allows evaluation of the practical relevance of the research outcomes. It provides the basis of power analyses and meta-analyses needed in future research” (p. xii). He went on to require that all articles submitted to the journal *Language Learning* include effect sizes. Although other journals are not currently requiring reports of effect sizes, the Fifth Edition of the *Publication Manual of the American Psychological Association* (2001) strongly encouraged the reporting of these statistics, emphasizing that “[f]or the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section” (p. 25).

Effect size is not dependent on sample size and therefore can allow comparisons (meta-analyses) across a range of different studies with different sample sizes. A standard measure of effect size is Cohen’s d (see app. J for the formula), which can be used to test differences in means between two groups or differences in gain scores between two groups. A value of .2 is generally considered a small effect size, .5 a medium effect size, and .8 or more a large effect size. Effect size can be calculated based on a number of statistics (correlations, parametric, and nonparametric). A useful reference is Wilkinson et al. (1999).

9.11. META-ANALYSES

There are times when our research questions involve surveying a wide range of existing studies rather than collecting original data. In most instances it will be difficult to directly compare studies given the unevenness of available data, size of experimental and control groups, and so forth. To make a meaningful comparison, effect sizes become the main comparative tool. Norris and Ortega (2000) exemplified this in a study on the effectiveness of second language instruction. They outlined the following five uses of effect sizes in their study:

- Average effect sizes across studies were calculated for specific predetermined categories.
- Average pretest to posttest effect sizes were calculated.
- Average effect sizes across studies were calculated based on duration of treatment.
- Average effect sizes were calculated for delayed posttests.
- Average effect sizes were calculated by type of dependent variable.

Thus, effect sizes can be a useful tool for researchers who want to compare results with other research that addresses similar questions.

9.12. CORRELATION

One differentiating factor between correlational research and what we have discussed in previous sections is that in correlational research no variables are manipulated. Correlational research attempts to determine the relationship between or among variables; it does not determine causation. Consider the fictitious example in the following box:

The Relationship Between Infant-Directed Speech and Growth Spurts

Introduction

A research team believes that talking to young children (infants) is related to their growth; the more talk addressed to young children, the more they grow. To test this, the team considers two mother/child pairs. They gather speech and growth data from children aged 6 months to 18 months (twice a month, 30 minutes each time). To measure the amount of talk, the researchers count all words in that 2-hour period. The following table shows the data for both mother/child pairs:

<i>Month of Data Collection (week)</i>	<i>Pair #1</i>		<i>Pair #2</i>		
	<i>Number of Words</i>	<i>Height in Inches</i>	<i>Month (week)</i>	<i>Number of Words</i>	<i>Height in Inches</i>
1 (1)	72	24	1 (1)	65	28
1 (3)	75	24	1 (3)	70	28
2 (1)	75	25	2 (1)	66	28
3 (3)	70	25.5	3 (3)	72	29
3 (1)	90	25.5	3 (1)	59	29.5
2 (3)	92	25.5	2 (3)	64	29.75
4 (1)	89	26	4 (1)	64	30
4 (3)	90	27	4 (3)	80	30
5 (1)	91	27	5 (1)	82	30
5 (3)	102	27.5	5 (3)	100	30

6 (1)	93	28	6 (1)	125	30.25
6 (3)	94	28	6 (3)	152	30.5
7 (1)	91	28	7 (1)	145	30.5
7 (3)	121	28.5	7 (3)	150	30.5
8 (1)	132	29	8 (1)	145	31
8 (3)	120	29.5	8 (3)	180	31
9 (1)	145	30	9 (1)	92	31.25
9 (3)	145	30	9 (3)	165	32
10 (1)	120	30	10 (1)	172	33
10 (3)	105	31.5	10 (3)	170	33.5
11 (1)	75	31.5	11 (1)	200	34
11 (3)	105	32	11 (3)	180	35
12 (1)	190	32	12 (1)	178	36

Problematic Interpretation:

Using the data presented in this table, a correlation coefficient of .70 for pair 1 ($p < .001$) and .82 (Pair 2) is obtained ($p < .001$). The team concludes that because there is a relatively high correlation this proves that the number of words used was the source of the growth.

What is wrong with this picture? There are a number of issues that could be raised but the important one for this section is the interpretation. The first is that nothing has been "proven." All that has been shown is that there is a relationship between the number of words used when addressing a child and a child's height. The relationship is not necessarily one of cause and effect. Although it is true that there is a relationship, the source of each variable is different. Increased amount of talk to an infant is possibly due to a relatively higher interactive capability of the infant; increased height is a natural part of increased age in most children.

This example focused on the interpretation of correlational data. We now turn to how to determine the strength of a correlation. Correlations are calculated between two sets of scores (in the previous example, one score is the amount of talk and the second is height). We can plot this information on a graph. The amount of talk can be plotted along the X-axis and the height on the Y-axis. The result would be a graph with many individual

points. If there were a relationship between the two scores, the dots would cluster around an imaginary line. When we calculate a correlation, we come up with a correlation coefficient (r) that characterizes the direction of the line and how well the line represents the patterns in the data. Depending on the direction of the line, correlation coefficients can be expressed as positive and negative values. A positive value means that there is a positive relationship; for example, the more talk, the taller the child. Conversely, a negative value means a negative relationship—the more talk, the shorter the child. A value of zero means that there is no relationship between the variables. These three possibilities are illustrated in Figs. 9.8–9.10. The first figure comes from the data showing a positive relationship between amount of talk and height of child for Pair 2. Figure 9.8 represents a graph that would depict a negative relationship, and Fig. 9.9 is a graph showing no relationship between two variables.

9.12.1. Pearson Product-Moment Correlation

We now turn to the Pearson product-moment correlation, a common means for determining the strength of relations (see its formula in app. J). There are four assumptions that underlie this particular statistic:

- Normal distribution.
- Independence of samples.

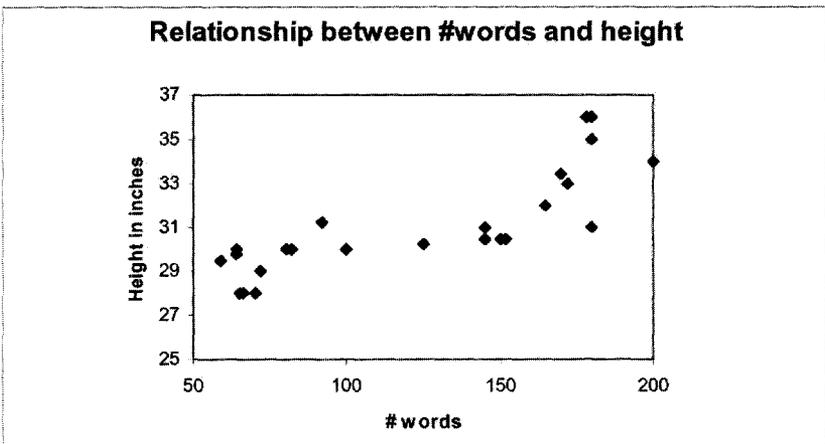


FIG. 9.8. Positive relationship ($r = .82, p < .001$).

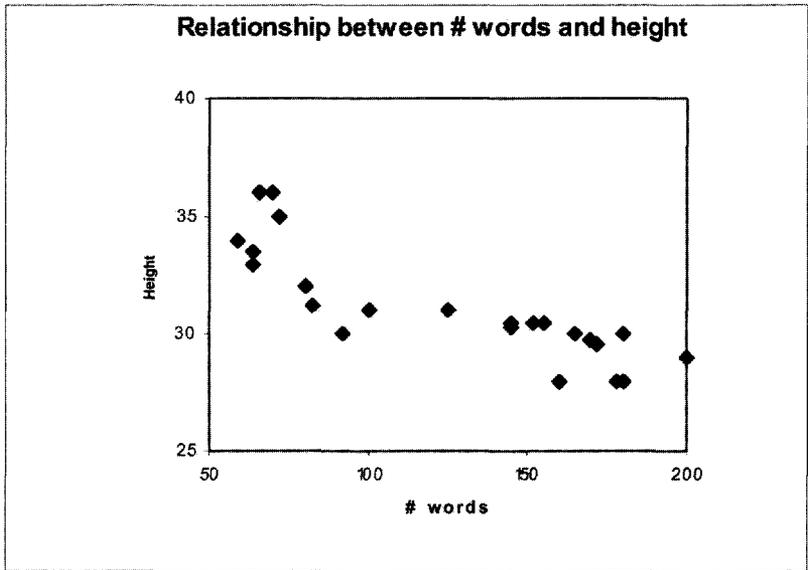


FIG. 9.9. Negative relationship ($r = -.85$, $p = .001$).

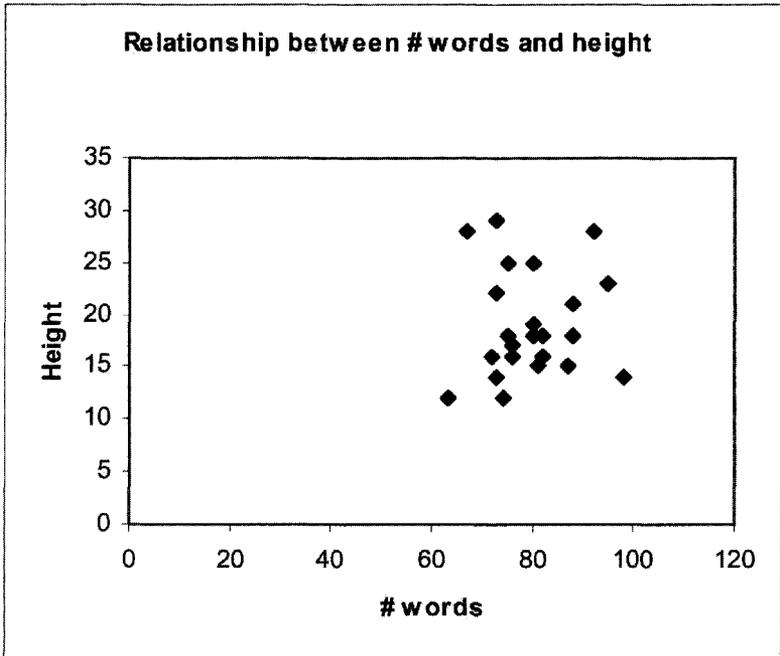


FIG. 9.10. No relationship ($r = -.051$, $p = .808$).

- Continuous measurement scale (generally interval or sometimes ordinal if continuous).
- Linear relationship between scores for each variable.

The correlation coefficient (which ranges from +1 to -1) gives information about the extent to which there is a linear relationship between the variables.

Frequently, correlations are calculated between multiple sets of scores in research studies. One concise way of presenting this data is in a correlation table, in which correlation coefficients for different sets of scores are listed. An example of a Pearson's correlation table (Table 9.10) comes from a study

TABLE 9.10
Pearson's Product Correlation Table: Correlations (Pearson's *R*)
Between the Language Aptitude Mean Scores
and the Mean Scores per Task Type and Test Session,
Under Explicit and Implicit Conditions

<i>Task Type</i>	<i>Test Session</i>	<i>Explicit (n = 27)</i>	<i>Implicit (n = 27)</i>
1	T1	.38	.15
	T2	.47*	.42*
	T3	.50*	.55*
2	T1	.21	.02
	T2	.34	.34
	T3	.39	.34
3	T1	.52*	.32
	T2	.56*	.50*
	T3	.54*	.39
4	T1	.19	.36
	T2	.45*	.51*
	T3	.40	.50*

Note: Task Type 1 = judgment task with time pressure; Task Type 2 = judgment task without time pressure; Task Type 3 = gap-filling task; Task Type 4 = correction task. T1 = midtest; T2 = immediate posttest; T3 = delayed posttest. $p < .01$

Source: de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19, 263. Copyright © 1997 by Cambridge University Press. Reprinted with the permission of Cambridge University Press.

by de Graaff (1997) on the role of explicit instruction versus implicit instruction in an artificial language by native speakers of Dutch as it relates to language aptitude. This table is to be interpreted in such a way that if we look at T2 in Task Type 3 in the explicit condition, there is a .56 correlation between the mean aptitude score and the immediate posttest score on the gap-filling task. The probability level is based on the value of the correlation coefficient and the sample size.

9.12.1.1. Linear Regression

We now turn to another use of correlations, that of prediction. Again considering our fictitious study of the relationship between infant-directed speech and growth, we repeat Fig. 9.8 below. As can be seen from the figure and the correlation coefficient (.71), there is a positive relationship, but if we had reason to believe that the relationship was meaningful, we might want to make predictions. For example, if the amount of words addressed to one specific child was 145, what might we expect his or her height to be? A straight line, called a regression line, might help us to address this question. A prediction equation can be used once we know the slope of the line and the intercept. Although the details of these calculations go beyond the scope of this chapter, it is useful to know that if we want to predict one variable from another, and we know details of the regression line, we can calculate, for any given words addressed, the predicted height. Note that the validity of regression for prediction is dependent on the variables selected.

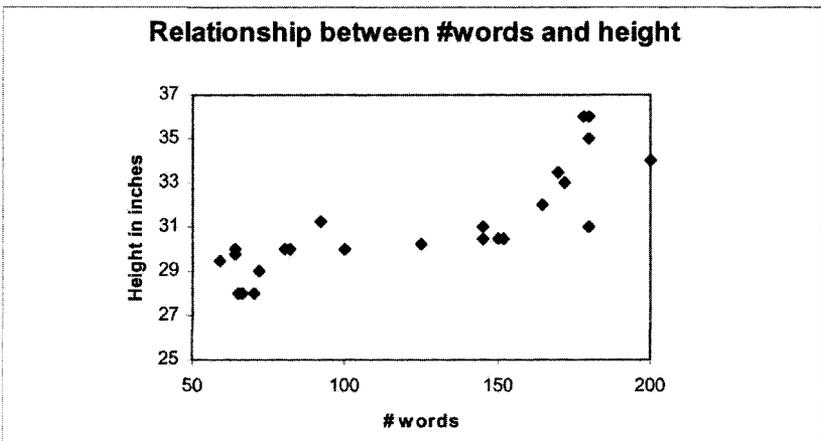


FIG. 9.8. (repeated)

A theoretically sound explanation for suspecting a relationship between the variables should be presented when regression is used to predict values.

9.12.1.2. Multiple Regression

There may be instances when we want two or more variables to be used to predict a third variable. We can use multiple regression to do that. For example, we might want to predict how ESL learners will do in college based on two factors: their results on a standardized test (e.g., TOEFL, [Test of English as a Foreign Language], and their performance in an Intensive English Program (IEP) on their own campus. A multiple regression prediction formula enables us to do this.

To test the validity of the predictor variables for predicting the third variable, data on the third variable should be collected from a subset of the population of study. For example, if we want to know how well TOEFL scores and IEP grades predict college grades, we could obtain actual college grades from a group of students and correlate our predicted grades based on a multiple regression formula with the actual grades. The resulting coefficient is called a coefficient of multiple correlation (R). The same idea applies as to r : R refers to the strength of the relationship among the variables in question (including the variable that is being predicted). Thus, as with other correlations, R can vary from $+1$ to -1 , with $+1$ being a perfect positive correlation and -1 a perfect inverse relationship. The higher the absolute R value, the more confident we can be in our predictions.

9.12.2. Spearman Rho/Kendall Tau

Both Spearman rho (r) and Kendall Tau are used for correlational analyses when there is ordinal data (or with interval data when converted to ranks). Spearman rho is more common, but the Kendall Tau is more suitable when there are many ties in the rankings.

9.12.3. Factor Analysis

Factor analysis is a complex procedure for determining common factors that underlie measures that test different (possibly related) variables. Researchers search for groups of variables that are correlated with one another; each group can be referred to as a factor. In doing a factor analysis, researchers take

into account the variance that is common to all individual variables, the variance that is specific to each variable, and sampling error. Factor analysis can be used to determine overall patterns found in correlation coefficients and is particularly useful when analyzing results from surveys.

In the preceding sections we have dealt with correlational research. In the next sections we deal with statistical packages that can assist in analyzing data and preparing those data for presentation.

9.13. STATISTICAL PACKAGES

There are commercially available statistical packages that can be used with either Macintoshes or PCs. In addition to the ones we address in this section, some basic statistics and graphing can be done using Excel. Two of the most common packages are SPSS and VARBRUL.³ Learning to use either of these requires some initial effort and practice, but there are many courses or workshops in the use of these programs. Specific details on statistical packages is beyond the scope of this book; but in the sections that follow we give an indication of the use that can be made of each.

9.13.1. SPSS

SPSS is a basic analytic program. There are add-on packages for more sophisticated statistical use, but the standard statistical tests such as frequency statistics (chi square), t-tests, ANOVAs (with post-hoc tests), regression, correlations, and other more complex statistics are included. One can also convert raw data and output from SPSS to charts and graphs. Further details on SPSS are available through <http://www.spss.com>.

9.13.2. VARBRUL

VARBRUL (Pintzuk, 1988; Rand & Sankoff, 1990) is a statistical package that is designed for analyzing variation data. For example, Young (1991) investigated *-s* plural marking in English by Chinese speakers. He wanted to

³SAS is another available statistical package (www.sas.com). It is more often used in business and in the hard sciences than in second language research, perhaps because it is perceived to be less user friendly than SPSS, although it is used within the domain of language testing. SYSTAT (www.systat.com) is another program for statistical analysis and display. An example of its use in second language research is to investigate bilingual education.

know what the possible influence might be that would predict when the English plural form was used. He hypothesized 10 possible influences, including who the interlocutor was (Chinese or English), whether the noun was animate or inanimate, whether the noun was definite or not, and phonological surroundings. Each category could be divided into at least two levels. Because of the large number of factors and because many of the cells have zero, an ANOVA is not appropriate for this comparison. VARBRUL, on the other hand, is designed to handle data of this sort. Young and Bayley (1996) provided detailed instructions on how to conduct a VARBRUL analysis and how to use data from this program.

9.14. CONCLUSION

This chapter has provided an overview of some of the statistical techniques commonly used in second language research. In the final chapter, we provide guidelines on what to do as you complete your research and prepare your findings for presentations and/or publication.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Which of the following research topics would lend themselves to a correlational analysis? Justify your answer.
 - a. Attitudes toward target culture and success in language learning.
 - b. Feedback type and language learning.
 - c. Attention paid to form and success in vocabulary acquisition.
 - d. Number of times male and female learners respond to feedback.
2. You are in charge of student services in an intensive ESL program and notice that the students who are doing well in their reading class are not doing well in their writing course. This is surprising, but you want to make sure that there is a relationship before bringing it to the attention of the director. Below are the two sets of scores from the 14 students enrolled in the course. You determine that the Pearson product-moment correlation is the most appropriate and can be used because you have interval data. Calculate the correlation coefficient (it has been started for you) and determine if the re-

sults are worth bringing to the attention of the director. All scores have a maximum of 100.

Using the formula in Appendix J calculate the correlation between the two class scores (you will either need a calculator or you can use a program such as Excel). Given your results, would you notify the Director? Why or why not?

	<i>Reading Class</i>	<i>Writing Class</i>
Maria	92	72
Juan	85	85
Toshi	78	93
Yoon-Soon	61	51
Bob	25	32
Sachiko	87	62
Young-Ahn	67	78
Gunter	59	57
Angelika	84	72
Noriko	85	82
Jean-Marc	77	55
Antonio	62	77
Giovanna	88	87
Susana	87	88
	$\Sigma X = 1037$	$\Sigma Y = 1001$
	$\Sigma X^2 = 1075369$	$\Sigma Y^2 = 1002001$
	$\Sigma XY = 1038037$	

- The statistical table that follows was discussed in this chapter. You have just conducted a study and have compared two means. You had 13 degrees of freedom. What must your critical value be to claim that your results are significant at the .05 level? At the .01 level? If you

had 15 degrees of freedom, what would the critical value have to be to claim significance at the .001 level? At the .05 level?

Distribution of t

<i>p</i> (<i>df</i>)	.1	.05	.02	.01	.001
11	1.796	2.201	2.718	3.106	4.437
12	1.772	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073

4. Consider the following data from a study comparing two groups of second language learners: one who received feedback and one who received no feedback. Compare the means of these two groups using the formula given in Appendix J for a t-test (maxi-

	<i>Feedback Group</i> (<i>n</i> = 9)	<i>No Feedback</i> (<i>n</i> = 8)
Mean	42	29
Standard deviation (SD)	1.23	1.59
Standard error of the mean (SEM)		
Standard error of the difference (SED)		

mum score was 50). (Note that there are somewhat different formulae that are to be used with different group sizes. Consult a statistics book for the precise formula.)

The t value is _____. With _____ df, the results are (not) significant.

5. Assume that you have gathered data on use by English speakers of correct noun-adjective gender agreement in Spanish. You have four groups of learners ranging from first-semester to fourth-semester Spanish classes. You want to determine whether learners from different classes are more likely to have acquired gender agreement. You define acquisi-

<i>Number of Correct Instances</i>	<i>Acquired Gender Agreement</i>	<i>Not Acquired Gender Agreement</i>	<i>Total</i>
1st semester	5	25	30
2nd semester	10	22	32
3rd semester	25	4	29
4th semester	26	6	32
Total	66	57	123

tion as 90% suppliance in obligatory contexts. Based on this standard, you determine which learners have and which have not acquired gender agreement. The data can be seen in the following table.

Calculate the chi square value using the formula in Appendix J where f_e is

Distribution of Chi Square

<i>p (df)</i>	.1	.05	.02	.01	.001
1	2.71	3.48	5.41	6.64	10.83
2	4.60	5.99	7.82	9.21	13.82
3	6.25	7.82	9.84	11.34	16.27
4	7.78	9.49	11.67	13.28	18.46
5	9.24	11.07	13.39	15.09	20.52

the expected frequency and f_o is the observed frequency. Using the following table, determine if the results are significant and then write a summary statement about how to interpret the results.

6. Describe a study in which measures of central tendency would be the only necessary analysis.
7. A researcher is interested in whether there is a connection between native language and ESL reading. He administers a reading comprehension test to high intermediate-level learners in a community education program. Means and standard deviations were calculated for each L1 group and are displayed in the following ta-

Hypothetical Reading Comprehension Study

<i>L1 Group</i>	<i>Mean ($x/100$)</i>	<i>SD</i>
Spanish	75	8.5
French	78	2.6
Russian	67	5.4
Mandarin Chinese	60	7.3
Korean	61	11.7

ble. How can these data be represented in a figure? Sketch two possible figures. Sketch one that also represents the standard deviation.

8. What are outliers, and which measure of central tendency is the most sensitive to data from outliers?
9. Read three articles that use statistical analyses (select articles that differ in the statistics used). For each:
 - a. What statistical analyses were used?
 - b. Why were these statistical tests used?
 - c. Were data presented in tables, graphs, or both?
 - d. If tables were used, describe the information presented in them.
 - e. If graphs were used, interpret them.

Concluding and Reporting Research

In this chapter, we provide information about the final stages of research projects, including tips for drafting sections or chapters in which results are discussed, together with limitations and conclusions sections. We also consider issues such as the audience for the research to be reported. We conclude the chapter with a detailed checklist for researchers to consider when research is being prepared for submission for publication or presentation. Although much of this chapter focuses on the more prescriptive requirements of concluding and reporting quantitatively oriented research, we also include information about preparing qualitative reports.

10.1. THE IMPORTANCE OF REPORTING RESEARCH

As we discussed in the beginning of this book, the purpose of research is to discover answers to pertinent questions. In order for answers to be meaningful, they must be reported to an audience. If research findings are not reported and heard or read, even the most carefully executed and elucidating studies are essentially meaningless. Therefore, we consider reporting findings to an audience to be one of the most crucial elements in the process of second language research. Reports of research generally involve a clear description of the problem and the methodology, together with the results, the researchers' interpretations of the data based on their theoretical framework, and a conclusion. The nature of research reports often differs for quantitative and qualitative studies, as well as for split or combined method designs. In this chapter on the final stages of research, we build on many of the terms and concepts discussed in earlier chapters. First, we discuss the final stages in research reporting; namely, the discussion and con-

clusion sections. This follows on from the previous chapters on the earlier stages of research such as data elicitation, coding, and analysis. Then, we move on to a consideration of the whole research report, and the steps to be taken before publication.

10.2. THE FINAL STAGES IN REPORTING QUANTITATIVE RESEARCH

In quantitative research, the results and discussion of findings may be combined into one section or presented in separate sections. The sections can be combined if the research design is relatively simple and the implications of the analyses are straightforward. However, with a more complicated research design and/or results that are complex, separating the results and discussion may add clarity to the reporting. The results section should include a clear description of the data collected and the outcomes of any statistical procedures. The analyses are often organized in terms of the relevant research questions or hypotheses so that it is clear whether the hypotheses are confirmed or rejected. In discussion sections, the authors present their interpretation of the results. In addition, they should address the implications the results might have for theory and/or practice, together with the limitations of the study and any suggestions they may have for further research.

Although the structure and organization of the final sections of research reports can vary from author to author and from one type of study to another, these final sections do share several common elements in second (and first) language research. As can be seen in Table 10.1, the closing segments of articles in different areas of second language research typically include discussion sections, together with additional sections concerned with the limitations, conclusions, and sometimes the pedagogical implications of the research. We outline the common elements of each of these sections in more detail in the discussion that follows, and provide examples from the seven studies cited in Table 10.1, which represent a range of interests in second language research.

10.2.1. The Discussion

To get an overview of the information that is typically included in a discussion section, it is helpful to review the organization of the final sections of some sample articles from the field of SLA, presented in Table 10.1. Six of the seven articles include discussion sections; Willett (1995) did not dedi-

TABLE 10.1
Organization of Final Sections
in Language Research Journal Articles

<i>Berg (1999)</i>	<i>Ellis and He (1999)</i>	<i>Leow (2000)</i>	<i>Willett (1995)</i>	<i>Philp (2003)</i>	<i>Williams (1999)</i>	<i>Melzi and King (2003)</i>
experimental (quantitative) study of the effect of trained peer response on L2 writing	experimental (quantitative) study of modified output and L2 learning	experimental (quantitative & qualitative) study of the effects of awareness on intake	qualitative ethnographic study of L2 socialization	experimental (quantitative) study of recasts and L2 learning	classroom-based (quantitative) focus on form study	descriptive (quantitative) study on the use of Spanish diminutives in mother-child conversations
Discussion	Discussion	Discussion		Discussion	Discussion	Discussion
Future research		Limitations and future research		Limitations and future research	Pedagogical implications	
	Conclusion	Conclusion	Conclusion	Conclusion		Conclusion

cate a section exclusively to discussion, but instead chose to weave the discussion and results together. The other six discussion sections all included a summary of the results, an explanation of possible reasons for the results, a comparison of the results to those obtained in other studies, and a commentary on the significance or implications of the results. Each of these is discussed in more detail in this chapter.

Of these four elements, a summary of the results is the most common. This is often accomplished by referring to the specific research questions posed at the beginning of the study and providing succinct answers to those questions. Alternatively, the discussion can begin with a concise summary of the findings that were detailed in the results section and then move on to a more detailed discussion of each research question. Including such summarizing comments is useful in that it can clearly inform the reader, especially the reader who is skimming, as to the purpose and outcome of the study. Here are respective examples from the discussion sections of two of the studies outlined in Table 10.1:

- Ellis and He (1999): “The first research question asked about the relative effects of premodified input, interactionally modified input, and modified output on L2 learners’ comprehension. The results of this study indicate that reasonable levels of comprehension can be achieved in all three conditions” (p. 297).
- Philp (2003): “In general, the results support the claim that learners notice a considerable amount of implicit feedback provided through interaction in a primed context” (p. 114).

Another typical element in the discussion section is the provision of possible explanations for the results found in the study—that is, for researchers to go beyond merely reporting the results and speculate about why particular results were obtained. For example:

- Ellis and He (1999): “Why did the modified output group consistently outperform the two input groups in comprehension and vocabulary acquisition? ... We believe that the modified output condition afforded the learners a qualitatively different discourse experience” (pp. 297–298; they went on to support this explanation with examples and further details).
- Williams (1999): “It appears that learners, at least at lower levels of proficiency, do not frequently focus on formal aspects of language. One logical reason for this is that lower-level learners may have

enough to do just to maintain communications and they are therefore unable to focus on form to the same degree as the more proficient learners” (p. 612).

Comparing the results of the present study with the results found in earlier studies is also a common element in discussion sections. Earlier findings should have been presented in the literature review as a way of contextualizing the study. Comparisons of results in the discussion section have the effect of helping the reader understand how the findings relate to previous work. The findings may provide evidence, which supports and extends earlier findings, or they may indicate that existing frameworks need to be reconsidered and revised. For instance:

- Leow (2000): “[Although the findings] provide further empirical evidence for the association between awareness and subsequent processing of L2 data found in other classroom-based studies, [they do not support] a dissociation between awareness and learning as espoused by some researchers” (p. 568).
- Berg (1999): “Findings in this investigation lend support to the view often expressed in the literature that training is important for successful peer response” (p. 230).

Discussion sections also often provide comments about the significance of the results or their implications for either pedagogy or theory, which can make it easier for readers to incorporate the findings into a framework they already know. For an audience of researchers such comments may clarify fundamental concepts and facilitate further research, whereas an audience of teachers may be able to consider the results actively in relation to what they do in the classroom on a daily basis. The first two examples that follow point out pedagogical implications, whereas the third focuses on theory:

- Berg (1999): “Peer response can be an important learning tool in a writing course because it helps student writers do what they cannot yet do for themselves, and detect incongruities in their texts” (p. 232).
- Williams (1999): “It is possible to pinpoint activities that are more likely to foster such a focus [on form] than others Teachers can use this knowledge as they plan lessons and actively encourage students to engage in activities that draw or attract attention to form-meaning connections” (p. 619).

- Leow (2000): “From a theoretical perspective, no dissociation between awareness and learning was found in this study, the results of which are compatible with the claim that awareness plays a crucial role in subsequent processing of L2 data” (p. 573).

In sum, discussion sections are often central components of the final stages involved in reporting research, providing the important functions of summary, explanation, comparison, and appraisal. They can inform readers as to the purpose of the research, call attention to its context and implications, clarify theories and concepts, and promote further investigation and analysis.

10.2.2. Limitations, Future Research, and Conclusion Sections

Once research projects have been concluded, the provision of a full set of information about the limitations of a study is possible. This may be included in the discussion section; however, limitations may also appear as a separate section, or even as part of the conclusion. Regardless of the precise location, acknowledgment of the limitations of the research is important, not only as a caution to the readers against overgeneralization of the findings, but also as a suggestion for how future studies could be improved and as an indication of possible avenues for further investigation. As noted in chapter 1 and as can be seen in the following examples, such sections are often an important and rich source for research questions by other researchers:

- Willett (1995): “The question we must ask is not which interactional routines and strategies are correlated with successful language acquisition. Rather, we must first ask what meaning routines and strategies have in the local culture and how they enable learners to construct positive identities and relations and manage competing agendas” (p. 499).
- Berg (1999): “It is important to discover what takes place during trained versus untrained peer response negotiations ... [and] it would be useful to study the different aspects of the training program to determine the most useful activities for reaching desired outcomes” (p. 233).
- Leow (2000; suggesting that future studies should focus on methodology): “Robust research designs are clearly needed to address the issue of how representative participants’ performance in experimental

groups truly is, especially in the areas of attention and awareness in SLA” (p. 574).

- Melzi and King (2003): “It would also be of interest to collect and analyse similar data among younger children in order to investigate the developmental trajectory [of diminutive affix use]” (p. 302).

These sorts of comments help generate ideas for further research, point to areas that need to be addressed in order to advance the field, and, in general, serve as useful pointers for both novice and experienced researchers. In finalizing research for publication, the specification of directions for future research also demonstrates to the reader that the author has concluded one phase of the research and has carefully thought about the next phase.

Limitations sections usually also include a discussion of the generalizability of the results given the characteristics of the participants (e.g., L1, age, gender, proficiency level, socioeconomic status, instructional or experimental context for the research, country of origin, length of residence, etc.) or the linguistic focus of the study. For example, the first three of the studies in the following list imply limited generalizability and a corresponding need for replication, whereas the last implies that linguistic and cross-linguistic study are necessary:

- Philp (2003): “The majority of learners were educated to at least postsecondary level, most were socio-economically advantaged, and, in general, they were motivated to study the L2” (p. 118).
- Willett (1995; in relation to instructional context): “The kinds of interactional routines and strategies used in this particular classroom were local, not universal” (p. 499).
- Leow (2000): “The findings clearly cannot be extrapolated to other linguistic forms or structures” (p. 573).
- Melzi and King (2003): “[Our study] provides further support for the recent calls (e.g., Lieven, 1994) for increased study of non-English languages in the field” (p. 303).

Also to be found in limitations sections are contextual concerns (e.g., the need for a wider range of contexts to be addressed, including second vs. foreign language, or different discourse contexts), statements about the effect of the materials (e.g., tasks and tests) on the results obtained, the role of setting (e.g., different types of L2 classroom, communicative vs. forms oriented), the need for longitudinal studies investigating the long-term effect of the treatment used, and also interaction effects such as the relationship

between learner-internal factors, target language knowledge, and the effects of instruction and interaction.

To summarize, then, the preceding discussion delineated several common elements in the final sections of second language research articles:

- Summary of the results.
- Explanation of possible reasons for the results.
- Comparison of the results to those obtained in other studies.
- Commentary on the significance or implications of the results.
- Discussion of limitations.
- Suggested areas for further research.

It is important to remember that the exact sequence of these sections is not fixed, nor do all articles contain all elements. As Ruiyang and Allison (2003) noted, “The structure of empirical RAs [research articles] in applied linguistics tends to be flexible towards the end, partly because rhetorical functions can overlap” (p. 381). It is important to bear in mind when examining checklists such as these that many research reports—whether articles, grant proposals, grant reports, book chapters, or books—can be written in sections and checked in sections, just as we propose here. However, consistency among sections is crucial as well, so researchers also need to check for obvious contradictions and repetition of information, in addition to making certain that the sections logically match each other.

10.3. THE FINAL STAGES IN REPORTING QUALITATIVE RESEARCH

In the opening sections of reports, qualitative and quantitative research can be quite similar. For example, regardless of paradigm, authors usually clearly state their (initial) research questions and problems, and provide a theoretical background for their research. The context in which the research has been conducted should also be addressed, along with the many issues involved in the selection of participants. However, qualitative and quantitative research can be different in their final stages insofar as qualitative research reports can be more varied in terms of organization and in terms of the specific sections included (e.g., results and discussion might be grouped together). They also demand persuasive and skilled writing in order to effectively summarize large amounts of data and to communicate the significance of the research to the reader. Qualitative research must also

address many other elements that are inherent to non-quantitative research, as discussed next.

Different paradigms of qualitative research potentially involve distinct standards for reporting and stylistic elements. As we have discussed in chapters 3 and 6, qualitative research can involve a range of data collection methods, including, for instance, structured and unstructured classroom observations, structured and unstructured informal interviews, case studies, introspective analyses, and diary studies. Because it seems that acceptable reports vary based on the research paradigm and methods that the qualitative researcher adopts, qualitative researchers must decide how to organize their reports so that their ideas are best communicated to the intended audience.

Heath (1997) suggested that qualitative reports include introduction, research paradigm, and research method sections, and that they address preliminary biases, suppositions, and hypotheses. The introduction to qualitative reports might begin with a quotation or a vignette before describing the research question and situating it within a theoretical context. The research design section should be used to represent the epistemological, conceptual foundations and assumptions of the qualitative research paradigm chosen and should contain citations of authors who have defined the paradigm, thus increasing the validity of the design. The research methods section should include sufficient detail in order to increase its verisimilitude (i.e., authenticity and credibility). As such, the instrumentation used to collect the data, as well as the specific procedures followed, should be described. Reports should clearly state how the researcher gained access to participants and what kind of relationship was established between the researcher and participants. The nature of the data and how they were collected should also be clearly stated.

Particularly important for qualitative research is the inclusion of information about procedures such as how decision making was carried out and how the researcher implemented data reduction and reconstruction. It is also important for researchers to provide a clear sense of how much data were collected (e.g., how many interviews, and of what length, how many hours of observation, and over what period of time), because this is vital in assessing the strength of the research overall. Finally, as noted earlier, because researchers are usually primarily responsible for data collection and analysis, they need to report any preliminary biases, suppositions, and hypotheses prior to as the study, as well as whether and if so how these changed over the period of the study. For instance,

case study reports must be certain to specify the role of the researcher, because this is often more significant than that of a simple observer and may be relevant when interpreting the data. The boundaries of the case study must also be clearly described and motivated—for instance, why a particular case was selected, and how and in what contexts data were collected. In addition, although generalizations are seldom made based on case studies, the researcher should not only report findings but also draw conclusions that contribute to an overall understanding of a phenomenon within a theoretical framework.

Like case studies of individuals, classroom observation research should make the role of the researcher in the classroom explicit. If an observation instrument is utilized, a full description of it needs to be reported. On the other hand, for unstructured classroom observations, the research report might need to focus on how data were tracked as well as the decision-making process that led to the study's focus. If a data collection instrument was adapted or designed and revised, this should be made clear, often with much more detail about the processes that led to the revisions than would be the case for quantitative research. If surveys or questionnaires are employed to supplement and triangulate qualitative data, the researcher should report issues such as what the response rates were, whether or not there was a nonresponse bias, how analyses were performed, and whether any generalizations can be drawn from the results. It is also common to include copies of survey or interview questions in the appendixes. In summary, then, each qualitative research paradigm requires a unique consideration of its crucial elements when a report is written, in part because there are different research paradigms and many means of collecting and analyzing data. Because of this, researchers need to take particular care to detail (and justify) how they collected and analyzed their data.

When reporting the results of a qualitative study, researchers should also take into account the importance of rich or thick description. If the purpose of the research is to describe and classify the observed data, rich description is often utilized. The evidence reported should be detailed, multilayered, and comprehensive. Rather than reporting a limited number of anecdotes that support the conclusions, researchers should try to provide detail about a systematic selection of the data that represents both the central tendencies and variations. Researchers often also opt to present counterexamples as well: The purpose of some qualitative research, such as ethnographies, is to go beyond mere description and attribute observations to underlying constructs and systems of meaning.

This type of research will need to employ thick description, as discussed in chapter 6. An important question in qualitative research write-ups is how much interpretation of the data the writer should provide. Many qualitative researchers suggest that although writers may offer their own interpretations, they should also provide an adequate basis for their readers to construct their own independent interpretations. This may be accomplished by separating presentation of data (e.g., vignettes, interview excerpts, etc.) from discussion and analysis.

10.4. REPORTING COMBINED METHOD (QUANTITATIVE AND QUALITATIVE) RESEARCH

As discussed in chapter 6, it is becoming the case that quantitative and qualitative research methods are not viewed as dichotomous. Also, survey-based research methods, such as the use of questionnaires, are often used to triangulate both more quantitative and more qualitatively oriented data. However methods are classified, second language researchers are increasingly taking into account the fact that data can be collected using a wide range and combination of methods. When included in a primarily quantitative report, qualitative data or analytic techniques may provide unique insights that would escape both the researcher and the reader if statistical counts and analyses were used in isolation. For example, we have argued elsewhere (Gass & Mackey, 2000) that stimulated recall protocols, when collected and coded, often provide a particularly rich source of information that can elucidate a trend, exemplify any variation in the data, or provide insights into results that turn out to be different from what was predicted. Similarly, qualitative reports may become clearer when some quantitative analysis is included. Although a qualitative researcher may not be able to (or choose to) utilize parametric and comparative statistics, descriptive statistics can help make any tendencies or patterns in the data clear to readers. For example, graphs representing the data frequency distribution, measures of central tendencies (means, modes, or medians), and range and standard deviation characteristics of the data can help confirm the validity of any trends, patterns, or groupings that the researcher has identified through a qualitative analysis. Hence, it may be best if researchers, even if they identify their research as primarily qualitative or quantitative, not rule out the inclusion of both types of data in their reports. In reporting their studies, researchers need to consider all elements and requirements that will best explain the data to the audience.

10.5. CHECKLIST FOR COMPLETING REPORTS OF RESEARCH

When the author is ready to draft the final sections, it is useful to carefully evaluate the research before submitting it for publication or review. It is best if researchers can consider where they might want to submit research for publication before investing the time in drafting the final sections, because research may need to be written up differently depending on the target journal or publisher. Even for dissertations or masters theses, it can be helpful if the researcher has a publication goal in mind after completing the empirical work. Next, we provide a list of questions to consider when finalizing quantitative research.

10.5.1. The Research Problem and Questions

When reporting research, the problem and questions need to be clearly stated and presented as part of a theoretical framework. Helpful questions to ask include the following:

- Are the research questions motivated by the literature review/your discussion of the literature?
- Are the research questions clearly formulated and unambiguously worded?
- Are the research questions appropriate for the theoretical framework?
- Why is the central research problem worth investigating? Is the argument for why the study is interesting clearly presented? For example:
 - Does the study fill a gap in the literature by addressing a relatively underresearched area or an unresolved problem?
 - Does the study address a methodological concern observed in previous research?
 - Does the study replicate previous research? If it is a partial replication, is the new element clear and well motivated (e.g., a replication with a different population of learners, in a new context, or with different measures of learning)?
 - In general, how does addressing these research questions make an original contribution to the field?
- How have practical constraints that the researcher has faced—such as time, money, availability, and energy limits—impacted the investigation?
- Has the investigation of the research question avoided placing the participants in any physical or psychological danger? That is, are there ethical issues that should be considered/discussed?

10.5.2. The Research Hypotheses

The research hypotheses (if any) also need to be clearly stated and presented as part of a theoretical framework in the research report. It is important to note that not all quantitatively oriented papers specify hypotheses or predictions. If they are included, however, helpful questions to ask include the following:

- Are the hypotheses clearly stated?
- Do the hypotheses clearly specify the variables that might be related?
- Are the hypotheses appropriate for the theoretical framework?
- Are the hypotheses testable given the methods adopted for the research?
- Will the results lead to the generation of additional hypotheses to be tested in subsequent research?

10.5.3. The Audience

It is important to take into consideration the needs, interests, and expectations of the audience when reporting research. In choosing where to publish or present, researchers need to consider their audience and the match between their work and what is usually either published by the journal or press or presented at the conference venue. A useful place to start is to look at the publication venues of comparable articles that influenced the development of the research. With this in mind, useful questions include the following:

- Who is the primary audience for this article, report, paper, presentation, or book?
- Is there a secondary audience?
- Has the write-up been targeted to the relevant interests?
 - To ensure that the research is compatible with audience expectations, as well as to be sure that it is a good fit for a given journal or press, it is useful to skim a number of articles in recent issues of the journal, or some books recently published by the press.
 - For grant applications, it is also helpful to read the brief reports or summaries of previously successful grantees. These are often available on the Internet or by writing to the grant-awarding body.

- Have the needs and expectations of the readers been carefully considered? For example:
 - If the paper is for a class, have the assignment guidelines been followed?
 - If the article is to be submitted to a journal or a press, have the guidelines on page length, formatting, and reporting been followed?
 - If the research is part of a grant application or report, have the guidelines set by the grant-awarding body been followed?

10.5.4. The Abstract

As noted in chapter 1, the abstract provides a brief overview, that readers will usually use to determine whether the study is relevant to their current interests and research needs. However, in addition to enticing a potential audience and convincing publishers, grant awarders, or conferences to accept the research, it is also important to write a good, representative abstract for retrieval purposes. A large number of online search databases typically include only abstracts, and these are catalogued and indexed so that interested parties can search for relevant topics. Thus, useful questions to ask when evaluating the abstract for research reports include the following:

- Does it provide a concise yet representative overview of the topic and aim of the research?
- Are the sample and materials/methods briefly described?
- Are the results of the study summarized, and is the relevance and importance of the study clear?

10.5.5. The Literature Review

The literature review explains the context for the research, together with details about the findings, strengths, and weaknesses of previous studies in the area. Helpful questions to evaluate the literature review in the concluding phases of research include the following:

- Are all relevant studies surveyed?
- Does the review provide an accurate and objective summary of the current state of the art and the theoretical framework of the study?
- Does the review present readers with enough background to understand how the study fits in with other research?

- Now that the study is concluded, are any organizational changes or new inclusions to the literature review necessary to better contextualize the discussion of the results?
- Is the literature review relevant; that is, are studies that are peripheral to or irrelevant to the research question excluded?

10.5.6. The Design of the Study

In reporting studies, researchers must try to include enough detail about the design to allow other researchers to replicate the study and to be able to understand and evaluate the validity of the results, based on the methods used. Helpful questions include the following:

- Is it clear that the research design (e.g., experimental, quasi-experimental, correlational, qualitative, split method, etc.) was appropriate given the theoretical framework, purpose, and research questions of the study?
- Are all of the terms clearly defined and operationalized, with examples wherever space permits?
- Are each of the variables clearly defined?
- Is the design explained in sufficient detail to permit replication wherever possible?

10.5.7. Logistics

In the final stages of research and before reporting, researchers should also carefully address practical issues with a series of checks. For example:

- Has the appropriate permission or consent from the participants and all other relevant bodies (e.g., school boards, guardians, teachers, parents) been appropriately checked and filed?
- Are the data from the study kept in a secure place?
- Have all identifying details been kept confidential in the report wherever possible?
- Was there a contingency plan for a problem or unforeseen event that arose? If so, is information about how this was solved conveyed in the research report to assist future researchers who might face the same problem?
- Did any problems interfere with the basic timeline for the completion of the study? If so, should they be reported in order to help others who collect data in the same context?

10.5.8. Participants

Important questions involving the study's participants include:

- Are sufficient biographical details on the participants provided so as to permit replication?
- At the same time, have confidentiality/anonymity issues been taken into account when this information was reported?
- Is information on the selection or assignment of participants to particular groups provided?

10.5.9. Data Gathering

Questions addressing how the data were gathered should also be considered. Some helpful questions include:

- Is it clear that the choice of sample (e.g., random, nonrandom, stratified random) was appropriate given the purpose of the study?
- Is it clear that the means for gathering data was appropriate for the research question?
- Was evidence of the validity and reliability of the instruments provided in the write-up?
- Is sufficient and detailed information provided about how, when, and where the data were gathered?
- Was the status of the researcher made explicit in the data-gathering process? (Was the researcher an observer? A participant? What, if any, was the relationship of the researcher to the participants?)

10.5.10. Data Analysis

Some questions to keep in mind when reporting on data analysis include:

10.5.10.1. Transcription

- Was transcription of the entire dataset necessary for coding, or was partial transcription acceptable? Is this clearly reported?
- Were reliability checks performed on transcriptions (and results reported)?
- Is it useful for the target readers if detailed information is provided on how many transcriptions and/or hours of data were used?

- Is information about transcription conventions necessary for the audience? If so, is this information provided in an appendix or notes to any examples used?

10.5.10.2. Coding Systems

Question to bear in mind when reporting coding include the following:

- Was coding of the entire database necessary, or was partial coding acceptable?
- Is this clearly reported?
- Were reliable coding guidelines available?
- Are coding categories clearly defined and examples of each coding category provided?

10.5.10.3. Interrater Reliability

When reporting on interrater reliability it can be useful to bear in mind the following questions:

- Was an interrater (or intrarater) reliability check necessary for the research? If so, is the method of assessing such reliability clearly reported?
- Is the rationale for choosing that particular method explained?
- Was the interrater reliability statistic considered to be sufficiently high? If not, is information about the sources of disagreement provided? Is information about missed (rather than disagreed-on) episodes provided?
- Is information about the coders provided?
- Are the rating/scoring guidelines necessary for the reader to understand the coding system? If so, are they provided?

10.5.10.4. Data Organization

When organizing and describing the data it can be useful to ask:

- Do all important constructs in the research have clear theoretical definitions? Are all variables operationally defined?
- Is it clear that the constructs and variables were appropriate for the research?

- Is information provided clarifying the research variables and the scales that represent them?
- Is enough information provided for readers to determine what kinds of scales have been used for each variable?

10.5.10.5. Statistics

When reporting statistical tests it can be useful to ask:

- Are the statistical tests and procedures used clearly identified?
- If the choice of a statistical test is not one of the standard ones used in the field, is its appropriateness clearly demonstrated (e.g., by referencing statistics texts or published studies with similar data and tests)?
- If consultants were used, are they thanked in the author's note (discussed later in this chapter)?

10.5.10.6. Presentation

When presenting the data in the results section, helpful questions include the following:

- Are the data clearly summarized and presented in the report (e.g., charts, appendixes, figures, etc.)?
- Are both descriptions and graphical representations of the data included where appropriate?
- Is the chosen method of presentation the clearest, most effective, and most elegant way of presenting the data?
- Is appropriate and consistent formatting followed throughout?

10.5.11. Conclusions

As discussed earlier in this chapter, there are several considerations to keep in mind when drafting the final sections of research reports:

- Are the results of the research succinctly summarized?
- Are the limitations of the research acknowledged?
- Are the implications of the results for either theory or pedagogy (or both) discussed?
- Are suggestions as to the direction of future research provided?

10.5.12. References

Most style guides require researchers to ensure that *all* citations in reports of research, (whether books, articles, chapters, etc.), and *only* those citations, are included in the correct format in the reference list. Also, all citations in the research report should be consistent with the same style guide. It is important to consider carefully how secondary sources will be cited (e.g., when to include the complete and original reference in cases when you cite someone who cited someone). To illustrate the variation in referencing, following is a list of the different referencing styles used in seven journals specializing in second language research:

Language Learning, Studies in Second Language Acquisition, The Modern Language Journal:

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47, 101–143.

Applied Linguistics:

Duanmu, S. 1995. 'Metrical and tonal phonology of compounds in two Chinese dialects.' *Language* 71/2: 225–259.

Tomlin, R.S. and V. Villa. 1994. 'Attention in cognitive science and second language acquisition.' *Studies in Second Language Acquisition* 16: 183–204.

Second Language Research:

Spada, N. and Lightbown, P.M. 1993: Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition* 15, 205–224.

Language Learning & Technology:

Tomasello, M. and C. Herron. 1988. 'Down the garden path: inducing and correcting overgeneralization errors in the foreign language classroom'. *Applied Psycholinguistics* 9: 237–246.

System:

R. C. Gardner, P. F. Tremblay and A. Masgoret, Towards a full model of second language learning: an empirical investigation. *The Modern Language Journal* 81 (1997), pp. 344–362.

The style guide also needs to be followed for details such as the ordering of references in the text. Some guides require multiple citations to be in alphabeti-

cal order; others allow ordering to be chronological or selected by the researcher, perhaps in order of relevance to the point being made. Guidelines as to how to cite multiple publications by the same author in the same year—for example, Smith (1994a, 1994b)—also need to be checked with the style guide used. Meticulous checking of references and reference formatting, especially in cases in which multiple revisions have been made to a document, can be one of the most tedious and time-consuming factors involved in preparing reports of research. Software programs such as EndNote¹ (ISI ResearchSoft) have been designed to automate this task. In terms of the content, it is also obviously the responsibility of the author to ensure that the references are appropriate for the study described. The selection of sources is usually seen as related to the issues involved in writing the literature review and statement of the problem.

10.5.13. Footnotes, Endnotes, Figures, and Tables

Like the reference list, footnotes, endnotes, figures, and tables must adhere to a standardized presentation that is consistent with a single style guide. Whereas for theses and dissertations they are customarily placed at appropriate points throughout the text (footnotes), journals and other publications often require that they appear at the end of the research article (endnotes).

10.5.13.1. Footnotes and Endnotes

Because many style guides suggest citing authors parenthetically within the text itself, footnotes and endnotes are generally not used for citing sources; rather, they are used to include information that, although relevant, does not fit into the flow of the text. They may include supplemental content supporting an idea expounded on in the text, concessions to a contrasting point of view, additional sources for further reading on the topic, or copyright information. It is generally advised that footnotes containing supplemental content should explain only one basic tangential idea. If they are longer or more involved than this, style guides suggest that authors consider incorporating the information into the body of the paper or including it as an appendix. Both footnotes and endnotes are usually indicated and referenced with superscript Arabic numerals, consecutively numbered.

¹Endnote (www.endnote.com) is a software program designed to organize bibliographic references and place them in an appropriate format for the journal for which one is preparing an article.

10.5.13.2. Figures

Figures are used to display information discussed in the text in a concise format that is easy to comprehend. They generally consist of graphs, diagrams, charts, illustrations, or photographs. Second language researchers commonly use figures when an image would make an arrangement or relationship easier to visualize, or when a pattern of results would be clearer in visual format to augment or supplement a list of numbers in a table. A general rule of thumb is that a figure should not be used simply to duplicate textual information, nor as extra material, but rather as a helpful complement to or amplification of what is expressed verbally in the text. Stylistically, each figure should be referred to in the body of the paper, where the author should also indicate what in the figure is relevant to the issue under discussion. Also, each figure must be identified by a number, ordered consecutively, and given a brief and descriptive caption. It is often the case with most theses or dissertations, in which figures appear in the text itself, that the figures are also listed in a separate table. Chapter 9 provides examples of a range of typical figures used in second language research.

10.5.13.3. Tables

Tables are often used to present quantitative data, statistics, and analyses in a format that makes them easy to understand and facilitates comparisons. In writing research reports, it is important for researchers to decide what in their data is relevant and worthy of notice rather than to present every aspect of a large dataset. When discussing the data in a table, researchers should guide the audience toward the information they feel is significant. As with figures, tables must be identified numerically, numbered consecutively, and referred to in the text. They should be given brief, descriptive titles, and notes can be added to the bottom of the tables to explain specific aspects of their content. Even though they are discussed in the text, tables should also be relatively free-standing; any abbreviations and units of measurement should be defined, and all of the rows and columns must be appropriately labeled. As mentioned earlier, tables are placed at the end of articles submitted for publication. In the text of a thesis or dissertation, however, it is important that they be placed as close as possible to where they are mentioned in the body of the paper.

10.5.14. Author's Note/Acknowledgments

Many reports of research include a section in which people who have helped with the research are thanked. These notes appear in different sec-

tions, depending on the format of the report. For example, in journal articles and book chapters, an author's note of thanks may appear either as a footnote or as an endnote, whereas in a book it may be part of the preface or even the dedication. Researchers often wish to thank (a) their participants—for example, learners, native speakers, teachers, and so on; (b) any colleagues who may have read earlier drafts of the work and offered suggestions or feedback; (c) any assistants—including students, colleagues, or co-workers—who may have helped with data collection, materials development, transcribing, coding, or library work, etc.; and (d) any consultants who may have helped with statistics or ideas. Grant support is also often acknowledged in the author's note, usually with the grant number included. Anonymous reviewers are often thanked as well. Many authors finish their notes by stating that despite the help they have received with their articles, they are solely responsible for the content and for any errors.

10.5.15. Postresearch Concerns

Once the research has been completed, there are still tasks about which the researcher must think. For example:

- What feedback, if any, needs to be provided to the participants?
 - Do the participants need to receive brief, accessible reports about the research findings?
 - Do any other parties (e.g., teachers or administrators) need to receive such reports?
- If necessary, have any participants been given an oral debriefing on the purpose and outcome of the study? (Confidentiality must be considered when summarizing results for participants and involved parties.)
- Have the participants been thanked and/or remunerated? In the report or the author's note, have those who contributed to the research been appropriately acknowledged?

10.5.16. Final Touches and Formatting

When writing reports in a manner appropriate to the research paradigm, in addition to the content and organization of what is being reported it is important to consider how the report is presented. For example, the front and back material (title page, abstract, author note, references, and appendixes) and the formatting can be very important to publication. This material will differ depending on where the report is to be submitted for publication. If

the research is to be submitted for conference presentation or poster display, clear guidelines are usually provided in the calls for papers. As discussed previously, if the report is to be submitted to a journal in the field of second language research, many (although by no means all) journals follow the guidelines set forth in the *Publication Manual of the American Psychological Association*. Researchers should carefully consult the relevant style manual when preparing to submit their manuscript. These manuals contain clear guidelines as to how the manuscripts are to be formatted. For example, as mentioned earlier, tables and figures are often treated as back matter and not included in the text. The 2001 APA style manual specifies that tables should not include vertical lines to separate columns and that pages with figures should be numbered and titled on the back side so that the figure itself is “photo ready” (2001). Style guides also often suggest that researchers practice gender-neutral writing wherever possible, for example, by using the term *the learner*.

As far as research proposals are concerned, when soliciting grants, researchers must pay close attention to the format requirements laid out by the relevant agency. Although compelling content and arguments for the value of the research should be the most important component of a grant proposal, grant submitters must also adhere to the regulations for margins, font size, spacing, page length, and so on. Often a grant awarding body will have an office or web-based tool that evaluates proposals in order to ascertain their compliance with formatting regulations before the proposal is sent out to be reviewed. Researchers should pay attention to the front and back material that needs to be included ahead of time, because scrambling to complete these at the last minute can lead to problems. For example, a grant-awarding institution may require the inclusion of such items as customized (and often abbreviated) curriculum vitae or biographical statements, a timeline for the research, and usually a budget proposal together with a prose justification for the budget. All of these can be used by reviewers when determining whether the researcher has an appropriate research background, realistic expectations, and a feasible financial plan and timeline for implementing the proposed research.

Equally important in reporting research are the guidelines outlined by particular universities for theses and dissertations. For example, some universities require that students follow a style manual approved by the mentor regarding formatting issues such as quotations, footnotes, and other stylistic details. Much of the front and back material, however, may also need to follow specific university guidelines. Among other elements, guidelines are of

ten required in regard to the layout of the title page, the table of contents, the type of paper used, and the spacing and margins. Other common specifications include information about the formatting of the abstract; for example, for many schools, the abstract must contain a statement of the problem, the procedure or method followed, the results, and the conclusion; there is often a word limit. Stories of excellent dissertations that were sent back for revision by universities because of formatting violations often make the rounds of graduate schools, together with stories about how reformatting then held up graduation. In short, the “packaging” of a research report is crucial, whether it is being submitted to a university committee, a grant-awarding agency, a journal for publication, or a publisher for a book.

In summary, the final stages of research consist of creating a complete product. Not only must ideas, theories, results, and conclusions of quantitative and qualitative research be clearly communicated, but the professionalism of the researcher should also be demonstrated through careful attention to the formatting and presentation of his or her work. Also, the earlier in one’s research career one gets used to conforming to the “packaging” of research, the easier one’s writing life will be.

10.6. CONCLUSION

We hope this chapter has provided a useful point from which both novice and more experienced researchers can evaluate their research at the point of its conclusion and, ultimately, finalize studies that will make a significant and lasting contribution to second language research.

FOLLOW-UP QUESTIONS AND ACTIVITIES

1. Find two articles in which results and discussion sections are combined. Is this combination justified in each?
2. Consider the four elements commonly discussed as part of limitations sections: a summary of the results, an explanation of possible reasons for the results, a comparison of the results to those obtained in other studies, and a commentary on the significance or implications of the results. For the two articles you found for Question 1, discuss their limitations sections in the context of the four common elements.
3. Provide three guidelines for writing up and discussing the findings of qualitative research.

4. You have carried out a study of the relationship between scores on an aptitude test and the development of past time reference in English as a second language. However, rather than the 25 participants per group you had anticipated, you only have 6 per group, and your findings are not conclusive. What sorts of considerations should be borne in mind when deciding where to report the results of this research?
5. If research reports are to be submitted to a journal or a publisher for consideration for publication, how can the relevant style guidelines be obtained?
6. In the final phases of research, when and how should findings be communicated to the participants?

Appendix A
SAMPLE SHORT FORM WRITTEN CONSENT DOCUMENT
FOR SUBJECTS WHO DO NOT SPEAK ENGLISH

This document should be written in a language understandable to the subject.

CONSENT TO PARTICIPATE IN RESEARCH

You are being asked to participate in a research study.

Before you agree, you must be informed about (i) the purposes, nature, and time range for the research; (ii) any procedures which are experimental; (iii) any reasonably foreseeable risks and benefits to the research; (iv) any potentially beneficial alternative procedures; and (v) how confidentiality will be maintained.

Where applicable, you must also be told about (i) any available compensation should injury occur; (ii) the possibility of unforeseeable risks; (iii) circumstances in which the investigator may decide to halt your participation; (iv) any costs to you; (v) what happens if you decide to stop participating; (vi) when you will be told about new findings which may affect your willingness to participate; and (vii) how many people will be in the study.

If you agree to participate, you must be given a signed copy of this document and a written summary of the research.

You may contact name at phone number any time you have questions about the research.

You may contact name at phone number if you have questions about your rights as a research subject or what to do if you are injured.

Your participation in this research is voluntary, and you will not be penalized or lose any benefits if you refuse to participate or decide to stop.

Signing this document means that the research study, including the above information, has been described to you orally, and that you voluntarily agree to participate.

signature of participant date

signature of witness date

Appendix B
SAMPLE CONSENT FORM FOR A STUDY IN A FOREIGN LANGUAGE CONTEXT

Consent to Participate in Research

Project Name Conversation and second language development
Investigator _____ **Telephone** _____ **Email** _____
Sponsor _____

The University Institutional Review Board has given approval for this research project. For information on your rights as a research subject, contact _____

Introduction

You are invited to consider participating in this research study. We will be evaluating the effect of carrying out different activities on learning English as a foreign language (EFL). This form will describe the purpose and nature of the study and your rights as a participant in the study. The decision to participate or not is yours. If you decide to participate, please sign and date the last line of this form.

Explanation of the study

We will be looking at how different kinds of speaking activities help EFL learners in X country develop skills such as fluency and creative thinking. In particular, we are interested in the difference between activities that are carried out several times and activities that are done only once. We are also interested in comparing the language used during conversation activities with a native speaker with speaking activities done individually. About 105 students enrolled in xxx will participate in this study. You will carry out speaking activities with a native English speaker outside of class time on three different days. Each speaking activity will take approximately 15 minutes to complete. As part of the study, you will also complete some written practice activities, and do some individual speaking activities in the language lab during the regularly scheduled class times. Each quiz and individual speaking activity will take about 20 minutes to do. A tape-recorder will be used to record what you are saying during all speaking activities. All the activities will be completed over a nine-week period.

Confidentiality

All of the information collected will be confidential and will only be used for research and teacher training purposes. This means that your identity will be anonymous, in other words, no one besides the researcher will know your name. Whenever data from this study are published, your name will not be used. The data will be stored in a computer, and only the researcher will have access to it.

Your participation

Participating in this study is strictly voluntary. That means you do not have to be a part of the study. Your decision to participate will in no way affect your grade in any class. You will participate in the same activities, but nothing you say or do will be used as part of the data. If at any point you change your mind and no longer want to participate, you can tell your teacher. You will not be paid for participating in this study. If you have any questions about the research, you can contact _____ by telephone at _____, by email _____, or in person at _____ office.

Investigator's statement

I have fully explained this study to the student. I have discussed the activities and have answered all of the questions that the student asked. If necessary, I have translated key terms and concepts in this form and explained them orally.

Signature of investigator _____ Date _____

Student's consent

I have read the information provided in this Informed Consent Form. All my questions were answered to my satisfaction. I voluntarily agree to participate in this study.

Your signature _____ Date _____

Note: In many cases, a University IRB or ethics committee will also require approval and review by the overseas institution.

Appendix C
SAMPLE CONSENT FORM FOR A CLASSROOM STUDY

Consent to Participate in Research

Project Title: Learner Uptake in the L2 Classroom

Researchers

Names
Email Addresses
Telephone Numbers
Office Addresses

The University Institutional Review Board has given approval for this research project. For information on your rights as a research subject, call the Institutional Review Board office at this number:

Introduction

We are currently undertaking a study to explore the effect of different variables in the language learning classroom. This form will describe the purpose and nature of the study. Please take whatever time you need to discuss the study with the researcher. The decision to participate or not is yours. If you do decide to participate, please sign and date the last line of this form.

Background and purpose of the study

We are particularly interested in the relationships among instructional materials designed for use in the language classroom, the ways that teachers use those materials, and the amount and nature of students' learning. We hope to use what we learn to improve the quality of language learning and teaching, and contribute to the growing body of knowledge in the area of language learning research.

Total number of participants

About 12 people will take part in this study.

General Plan

During the study, a tape recorder will be used to record your teacher's interactions with the students during 6 lessons. Instructional materials completed during class may also be used as part of the data. The lesson will follow the ____ school curriculum and be no different from other lessons during the term.

Length of Study

The study will last for 7 lessons.

Confidentiality

Every effort will be made to keep the data collected confidential. We will disclose personal information about you only if required to do so by the law. However, we cannot guarantee absolute confidentiality. Whenever data from this study are published, your name will not be used.

Data Security

If information about your participation in the study is stored in a computer, the computer will not be part of a network and only the researchers will have access to the data.

New Findings

If you would like us to, we will contact you to explain the results of our study after the study has been concluded.

Payment

You will not be paid for participating in this study.

Your rights as a participant

Your participation in this study is entirely voluntary. You have the right to leave the study at any time. Leaving the study will not result in any penalty or affect your relations with your teacher or XX school. Should you decide to leave the study, tell your teacher or

a researcher. You will still participate in class activities, but nothing you say or do which happens to be on the tapes, and nothing you submit to your teacher, will be used as part of the data.

Problems and questions

Email _____ or call _____ if you have any questions or problems.

Call the IRB Office at this number: _____ with any questions about your rights as a research subject.

Withdrawal by researcher

The researchers may stop the study or take you out of the study at any time should they judge that you are no longer at the appropriate level for the study, or for any other reason.

Researcher's Statement

I have fully explained this study to the participant. I have discussed the procedures and treatments and have answered all of the questions that the participant has asked.

Signature of researcher _____ Date _____

Participant's consent

I have read the information provided in this Informed Consent Form. All my questions were answered to my satisfaction. I voluntarily agree to participate in this study.

Your name _____

Your signature _____ Date _____

Appendix D
SAMPLE INSTITUTIONAL REVIEW BOARD APPLICATION
GEORGETOWN UNIVERSITY, FORM 1

Georgetown University Institutional Review Board (IRB-c)
**Application for Social and Behavioral
IRB-c Review (Form C-1)**

<p><i>Please complete this form and return it to the following address for processing:</i></p> <p>Institutional Review Board (IRB-c) Social & Behavioral Sciences ...</p>	<p style="text-align: center;">FOR IRB-c USE ONLY</p>
--	--

Section One: Application Information

Principal Investigator	
Department	
E-mail address	
Title (Student, faculty, or staff)	
Responsible Participant (<i>member of faculty or official or administrative unit</i>)	

Appendices -p. 5

Note. Downloaded August 3, 2004 from <http://ora.georgetown.edu/irb/irbC.htm>
Reprinted with permission.

Is this research for your thesis/dissertation?	Yes • No •
Title of Project	Purpose of Project (one or two sentences)

Consultants or co-investigators, if any	Department or Institution
Estimated duration of total project	
Estimated total number of subjects (including control subjects)	
Age range of subjects	
Sex of subjects	
Where will study be conducted?	
Source of subjects	

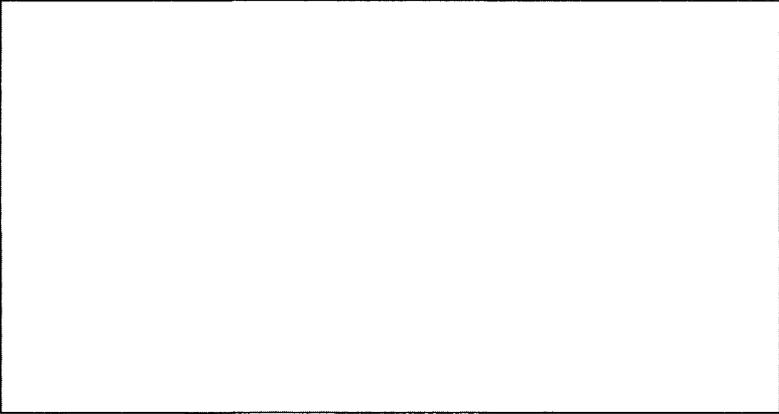
Grant Support for Project (if any)	Commercial Support (if any) for Project

Section Two: Information for IRB-c Review

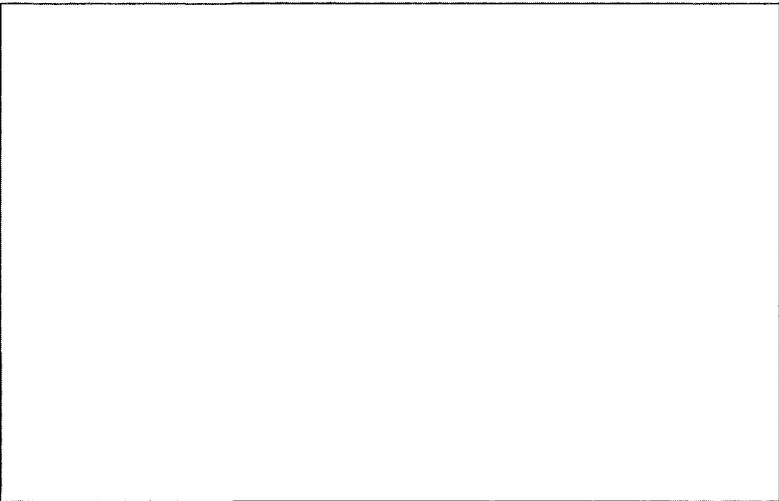
Please answer each specific question and use additional sheets as needed. A response of "See attached project description or grant application" is not sufficient.

- 1. Background.** Provide a brief historical background of the project with reference to the investigator's personal experience and to pertinent scientific literature. *Use additional sheets as needed.*

--



2. The plan of study. State the hypothesis or research question you intend to answer. Describe the research design, methods, interventions, and procedures (including standard or commonly used interventions or procedures) to be used in the research. Specifically, identify any interventions, procedures, or equipment that are innovative, unusual, or experimental. Where appropriate, provide statistical justification or power analysis for the number of subjects to be studied. *Use additional sheets as needed.*



3. Risks. Indicate what you consider to be the risks to subjects and indicate the precautions to be taken to minimize or eliminate these risks. If any data monitoring procedures are needed to ensure the safety of subjects, describe them. *Use additional sheets as needed.*

Section Three: Selection of Subjects and the Informed Consent Process

1. Populations. Indicate whether this project involves any of the following subject populations?

- Children (Children are defined by District of Columbia law as anyone under age 18.)
- Prisoners
- Pregnant women
- Cognitively impaired or mentally disabled subjects
- Economically or educationally disadvantaged subjects

If you indicated any of the above, in the space below, please describe what additional safeguards will be in place to protect these populations from coercion or undue influence to participate. *Use additional sheets as needed.*

2. Subjects. Describe how subjects will be recruited and how informed consent will be sought from subjects or from the subjects' legally authorized representative. If children are subjects, discuss whether their assent will be sought and how the permission of their parents will be obtained. *Use additional sheets as needed.*

3. Compensation. Will subjects receive any compensation for participation in cash or in kind?

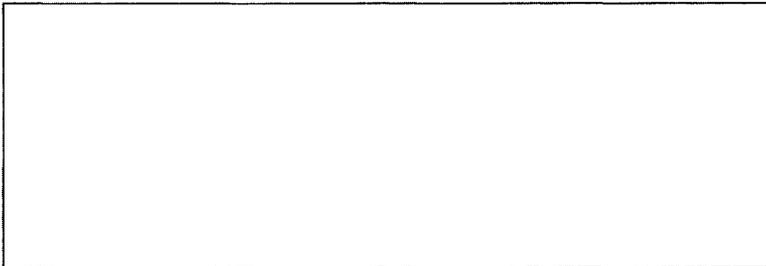
- Yes. *If so, please describe amount or kind of compensation in the space below.*
- No.

4. Fees. Will any finder's fees be paid to others?

- Yes. *If so, please describe the amount below.*
- No.

Section Four: Privacy and Confidentiality of Data and Records

1. Sensitive Information. Will identifiable, private, or sensitive information be obtained about the subjects or other living individuals? Whether or not such information is obtained, describe the provisions to protect the privacy of subjects and to maintain the confidentiality of data. Use additional sheets as needed.



Section Five: Conflict of Interest

1. **Conflict of Interest Form.** Do any investigators or co-investigators have a conflict of interest as defined in the Georgetown University Faculty Handbook? (See description on form C-6 at: <http://www.georgetown.edu/grad/IRB/irb-forms.html>)

- Yes. If so, please explain.
- No.

[Note: A copy of each investigator's and co-investigator's current Georgetown University Financial Conflicts of Interest Disclosure Form must be attached to this application (original plus one copy)]

2. Extramural Activity Form (Form C-5) required of all GU faculty/staff requesting IRD-C approval

I certify that the information furnished concerning the procedures to be taken for the protection of human subjects is correct. I will seek and obtain prior approval for any modification in the project design or informed consent document and will report promptly any unexpected or otherwise significant adverse effects encountered in the course of this study. I certify that all individuals named as consultants or co-investigators have agreed to participate in this study.

_____ Signature of Investigator	_____ Date
Department Chair: <input type="checkbox"/> Approved <input type="checkbox"/> Disapproved	
_____ Signature of Department Chair	_____ Date

If more than one department or administrative unit is participating in the research and/or if the facilities or support of another department are needed, then the chair or administrative official or each unit must also sign

_____	_____
Authorized Signature and Title	Date
_____	_____
Authorized Signature and Title	Date
_____	_____
Authorized Signature and Title	Date

Section Six: Attachments

Please attach the following items in order for the IRB-c to review your research:

Note: provide the original plus 15 copies of all materials for FULL REVIEW (ONLY). Provide the original only (of all materials) for EXEMPT and EXPEDITED reviews.

1. **IRB-c Application forms** (all forms are available on the IRB-c website at: <http://www.georgetown.edu/grad/IRB/>
 Form C-1 [always required]
 plus Form C-3 and/or Form C-4 (one or both, depending on nature of the research)
 plus Form C-5 (faculty/staff)
 plus Form C-6 [always required].
2. Certificate of completion of education in the protection of human research subjects (required).
3. The informed consent document.
4. Any recruitment notices or advertisements.
5. Any survey instruments, psychological tests (other than standard, commercially available instruments), interview forms, or scripts to be used in the research.
6. Investigator's qualifications (CV, biosketch, or Form 1572, if available).
7. Formal research protocol, if available.
8. Grant application, if applicable.

All IRB-c forms are to be submitted to the following address:

**Institutional Review Board (IRB-c) for the
 Social & Behavioral Sciences**

...

Appendix E
SAMPLE INSTITUTIONAL REVIEW BOARD APPLICATION
GEORGETOWN UNIVERSITY, FORM 2

Georgetown University Institutional Review Board (IRB-c)
Continuing Review Form (Form C-2)

<p><i>Please complete this form and return it to the following address for processing:</i></p> <p>Institutional Review Board (IRB-c) Social & Behavioral Sciences ...</p>	<p style="text-align: center;">FOR IRB-c USE ONLY</p>
---	--

Principal Investigator	
E-mail address	
Title of Project:	

1. What is the status of your research project?

- Active (still enrolling subjects)
- Closed to subject enrollment, but subjects still receiving or involved with research intervention
- All research interventions completed, but research open for data analysis and follow-up of subjects

Note. Downloaded August 3, 2004 from <http://ora.georgetown.edu/irb/irbC.htm>
Reprinted with permission.

Appendix F
SAMPLE INSTITUTIONAL REVIEW BOARD APPLICATION
GEORGETOWN UNIVERSITY, FORM 3

Georgetown University Institutional Review Board (IRB-c)
Request for Expedited Review (Form C-3)

<p><i>Please complete this form and return it to the following address for processing:</i></p> <p>Institutional Review Board (IRB-c) Social & Behavioral Sciences ...</p>	<p style="text-align: center;">FOR IRB-c USE ONLY</p>
---	--

Investigator:	Date:
E-mail address:	
Title of Project:	

Research activities that (1) present no more than minimal risk to human subjects and (2) involve only procedures listed in one or more of the categories below in Section One may be reviewed by the IRB through the expedited review procedure. Minimal risk means that the risks of harm anticipated in the proposed research are not greater, considering probability and magnitude, than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.

If you believe that your research falls into one of the following categories, please indicate which category or categories you believe is or are appropriate. One of the IRB Chairpersons or his or her designee will review your research to determine if expedited review is warranted. If warranted, your research will be reviewed to determine if approval can be granted. If granted, the form will be returned to you with an approval stamp in Section Three along with the signature of an IRB Chairperson, and you may begin your research. You must notify the IRB if your proposed research changes in any way. The IRB will request periodic updates. If expedited procedures cannot be used, the reason will be explained in Section Three, and your research must be reviewed during a convened IRB meeting.

Direct questions to the IRB Office at the address shown above.

Note. Downloaded August 3, 2004 from <http://ora.georgetown.edu/irb/irbC.htm>
 Reprinted with permission.

Section One: Categories Eligible for Expedited Review (*Please indicate one or more category, as appropriate, in the space next to the category numbers below.*)

- 1.* Research involving materials (data, documents, records, or specimens) that :
 ___ (a) have already been collected for some other purpose.
 OR
 ___ (b) will be collected for non-research purposes (such as medical treatment or diagnosis).
2. ___ Collection of data from voice, video, digital, or image recordings made for research purposes.
- 3.* Research on:
 ___ (a) individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior).
 OR
 ___ (b) research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.
4. ___ Continuing review of research previously approved by the convened IRB as follows:
 ___ (a) Where:
 (i) The research is permanently closed to the enrollment of new subjects, and
 (ii) All subjects have completed all research-related interventions, and
 (iii) The research remains active only for long-term follow-up of subjects,
 OR
 ___ (b) Where no subjects have been enrolled and no additional risks have been identified;
 OR
 ___ (c) Where the remaining research activities are limited to data analysis.
5. ___ Continuing review of research, not conducted under an investigational new drug application or investigational device exemption, where categories 2 through 8 do not apply, but the IRB has determined and documented at a convened meeting that the research involves no greater than minimal risk and no additional risks have been identified.
6. ___ Clinical studies of drugs and medical devices only when condition (a) or (b) is met:
 ___ (a) Research on drugs for which an investigational new drug application is not required. (Note: Research on marketed drugs that significantly increases the risks or decreases the acceptability of the risks associated with the use of the product is not eligible for expedited review.) **OR**
 ___ (b) Research on medical devices for which (i) an investigational device exemption application is not required or (ii) the medical device is cleared/approved for marketing and the medical device is being used in accordance with its cleared/approved labeling.
7. ___ Collection of blood samples by finger stick, heel stick, ear stick, or venipuncture from:
 ___ (a) Healthy, nonpregnant adults who weigh at least 110 pounds. For these subjects, the amounts drawn may not exceed 550 ml in an 8 week period and collection may not occur more frequently than 2 times per week;
 OR
 ___ (b) Other adults and children, considering the age, weight, and health of the subjects, the collection procedure, the amount of blood to be collected, and the frequency with which it will be collected. For these subjects, the amount drawn may not exceed the lesser of 50 ml or 3 ml per kg in an 8 week period and collection may not occur more frequently than 2 times per week.
- Note: 'Children' in (b) above is defined in the HHS regulations as "persons who have not attained the legal age for consent for treatments or procedures involved in the research, under the applicable law of the jurisdiction in which the research will be conducted" 145 CFR 46.402(a).*
8. ___ Prospective collection of biological specimens for research purposes by noninvasive means. Examples:

- (b) Deciduous teeth at time of exfoliation or if routine patient care indicates a need for extraction
- (c) Permanent teeth if routine patient care indicates a need for extraction
- (d) Excreta and external secretions (including sweat)
- (e) Uncannulated saliva collected either in an unstimulated fashion or stimulated by chewing gumbase or wax or by applying a dilute citric solution to the tongue
- (f) Placenta removed at delivery
- (g) Amniotic fluid obtained at the time of rupture of the membrane prior to or during labor
- (h) Supra- and subgingival dental plaque and calculus, provided the collection procedure is not more invasive than routine prophylactic scaling of the teeth and the process is accomplished in accordance with accepted prophylactic techniques
- (i) Mucosal and skin cells collected by buccal scraping or swab, skin swab, or mouth washings
- (j) Sputum collected after saline mist nebulization

9. ____ Collection of data through noninvasive procedures (not involving general anesthesia or sedation) routinely employed in clinical practice, excluding procedures involving X-rays or microwaves. Where medical devices are employed, they must be cleared/approved for marketing. (Studies intended to evaluate the safety and effectiveness of the medical device are not generally eligible for expedited review, including studies of cleared medical devices for new indications.) Examples:

- (a) Physical sensors that are applied either to the surface of the body or at a distance and do not involve input of significant amounts of energy into the subject or an invasion of the subject's privacy
- (b) Weighing or testing sensory acuity
- (c) Magnetic resonance imaging
- (d) Electrocardiography, electroencephalography, thermography, detection of naturally occurring radioactivity, electroretinography, ultrasound, diagnostic infrared imaging, doppler blood flow, and echocardiography
- (e) Moderate exercise, muscular strength testing, body composition assessment, and flexibility testing where appropriate given the age, weight, and health of the individual

* Note regarding categories 1 and 2: Some research in this category may be exempt from the HHS regulations for the protection of human subjects.

Signature of Investigator

Date

Section Two: Additional Materials

Please attach the following materials to this application:

1. IRB-c Application (Form C-1) (required)
2. Conflict of Interest (Form C-6) (required)
3. Informed Consent Form (if applicable)
4. Any survey tools or questionnaires (if applicable)
5. Certificate of completion of education in the protection of human research subjects (required)

Section Three: Committee Approval

FOR IRB-c USE ONLY

<input type="checkbox"/> Research Approved by Expedited Review (Category ____) <input type="checkbox"/> Expedited Review Not Allowed <hr/> Signature of IRB Chair or Designee Date	Comments:
--	-----------

Appendix G
SAMPLE INSTITUTIONAL REVIEW BOARD APPLICATION
GEORGETOWN UNIVERSITY, FORM 4

Georgetown University Institutional Review Board (IRB-c)
Request for Exemption (Form C-4)

<p><i>Please complete this form and return it to the following address for processing:</i></p> <p>Institutional Review Board (IRB-c) Social & Behavioral Sciences ...</p>	<p style="text-align: center;">FOR IRB-c USE ONLY</p>
--	--

Investigator:	Date:
E-mail Address:	
Title of Project:	

The human subjects regulations [45 CFR Part 46] define **research** as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge" [45 CFR 46.102(d)]. A **human subject** is "a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual or (2) identifiable private information" [45 CFR 46.102(f)].

However, some research involving human subjects may be exempt from the regulations. The categories below describe these exemptions. Please note that an exemption can be invoked only if **all** components of the research fit the category as described. You may find the following decision charts helpful:

<http://ohrp.osophs.dhhs.gov/humansubjects/guidance/decisioncharts.htm>

Note. Downloaded August 3, 2004 from <http://ora.georgetown.edu/irb/irbC.htm>
Reprinted with permission.

If you believe that your research may fall into one of the exempt categories, please indicate the relevant category in the space next to the category number below, and one of the IRB Chairpersons or his or her designee will review your research to determine if an exemption can be granted.

If granted, your exemption request will be returned to you with an approval in Section Three along with the signature of an IRB Chairperson, and you may begin your research. You must notify the IRB if your research changes in any way, because the exemption may no longer apply.

The IRB may request periodic follow-up. If an exemption cannot be granted, your exemption request will be returned to you with the reason listed in Section Three, and your research will be reviewed by the IRB. Please direct questions to the IRB Office at the address above.

Section One: Categories Eligible for Exemption *(Please indicate the relevant category in the space next to the category number. Categories continue on the next page.):*

NOTE: These exemptions do not apply to research involving prisoners, pregnant women, human fetuses, or human in vitro fertilization.

1. ____ Research conducted in established or commonly accepted educational settings, involving normal educational practices. Examples include:
 - a) Research on regular and special education instructional strategies,

OR

 - b) Research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

2. ____ Research involving the use of educational tests (cognitive, diagnostic, aptitude, or achievement), survey procedures, interview procedures, or observation of public behavior.

NOTE: Except as noted above, this exemption applies to all such research involving ADULT subjects unless BOTH of the following conditions apply:

- a) Information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects (NOTE: Codes constitute identifiers.);

AND

- b) Any disclosure of the subjects' responses outside of the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

NOTE: This exemption applies to research involving CHILDREN EXCEPT that (i) research involving survey or interview procedures with children is NOT EXEMPT, and (ii) research involving observation of the public behavior of children is NOT EXEMPT if the investigator(s) participate(s) in the actions being observed.

3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under category 2 above, IF:
 - a) The human subjects are elected or appointed public officials or candidates for public office;

OR

- ___ b) Federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens (**existing** means research materials are already on the shelf or archived when the research is proposed; e.g., blood samples already taken from patients or subjects for other clinical or research purposes). This exemption applies if:

- ___ a) These sources are publicly available,

OR

- ___ b) The information is recorded by the investigator in such a manner that individual subjects cannot be identified, directly or through identifiers linked to the subjects.

5. ___ Research and demonstration projects that are designed to study, evaluate, or otherwise examine:

- a) Public benefit or service programs;
- b) The procedures for obtaining benefits or services under such programs;
- c) Possible changes in or alternatives to such programs or procedures; or
- d) Possible changes in methods or levels of payment for benefits or services under such programs.

NOTE: This exemption applies ONLY to research and demonstration projects studying FEDERAL programs, and its use must be authorized by the Federal Agency supporting the research. As with all exemptions, IRBs and institutions retain the authority not to invoke the exemption, even if so authorized by the relevant Federal Agency. Studies of state and local public service programs require IRB review. Waiver of informed consent is possible for such programs under 45 CFR 46.116(c).

6. Taste and food quality evaluation and consumer acceptance studies, which meet any of the following conditions:

- ___ a) If wholesome foods without additives are consumed;

OR

- ___ If a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection of the US Department of Agriculture.

Signature of Investigator

Date

Section Two: Additional Materials

Please attach the following materials to this application:

1. IRB-c Protocol Application (Form C-1) (required)
2. Conflict of Interest Form (Form C-6) (required)
3. Informed Consent Document (if applicable)
4. Any survey tools or questionnaires (if applicable)
5. Certificate of completion of education in the protection of human research subjects (required)

Section Three: Approvals

FOR IRB-c USE ONLY

<input type="checkbox"/> Exemption Allowed (Category _____) <input type="checkbox"/> Exemption Not Allowed (Please see Comments) <hr/> <p>Signature of IRB Chair Date</p>	Comments:
--	-----------

Appendix H
SAMPLE TRANSCRIPTION CONVENTIONS
'JEFFERSONIAN' TRANSCRIPTION CONVENTIONS

1. [[Simultaneous utterances:
TOM: [[I used to smoke a lot when I was young

BOB: [[I used to smoke Camels
2. [Overlapping utterances:
TOM: I used to smoke a lot more than this
BOB: [I see
3.] End of overlapping or simultaneous utterance
TOM: I used to smoke a lot more than this.
BOB: [] Did you really?
4. = Linked or continuing utterance (no overlap)
 - a. for different speakers
TOM: I used to smoke a lot=

BOB: =He thinks he's real tough
 - b. for even more than one
TOM: I used to smoke a lot=

BOB: =[[He thinks he's real tough

ANN: =[[So did I
 - c. in either direction
TOM: I used to smoke a lot=
BOB: [I see=

ANN: =So did I
 - d. for the same speaker, simply to indicate that the turn continues
TOM: I used to smoke a lot more than this=
BOB: [You used to smoke-

TOM: =but I never inhaled.

5. Intervals

- a. (0.6) In tenths of a second
LIL: When I was (0.6) oh nine or ten
and between utterances
HAL: Step right up
(1.3)
HAL: I said step right up
- b. - Untimed, brief pauses
LIL: When I was - oh nine or ten
or longer
LIL: When I was - - - oh nine or ten
and between utterances
HAL: Step right up
((pause))
HAL: I said step right up

6. Delivery

- a. : Length
RON: What ha:ppened to you
and longer
RON: What ha::ppened to you
- b. . Falling ("final") intonation (followed by a noticeable pause)
HAL: Step right up.
- c. , Continuing ("list") intonation (slight rise/fall, followed by a short pause)
BOB: I saw Bill,
- d. ? Rising ("question") intonation (followed by a noticeable pause)
ANN: He left?
- e. underscore Emphasis
ANN: He left?
- f. (hhh) breathe out
(.hhh) breathe in
DON: (.hhh) Oh, thank you.

FRED: (hhh) That's a break.
- g. (()) Noises, kinds of talk
TOM: I used to ((cough)) smoke a lot
((telephone rings))
BOB: ((whispered)) I'll get it
- h. ! Animated talk (wider intonational contours)
BETTY: Look out for that rock!

- i. (h) 'Breathiness' (often laughter during speech)
TOM: I woul(h)dn't do that.
- j. x- Abrupt cutoff (glottal)
BETTY: Look ou-
- 7. () Transcriber doubt
TED: I (suppose I'm not)
(BEN): We all (t-)
ANN: ()
(spoke to Mark)
LIL: I
(suppose I'm not)
- 8. [] Phonetic transcription
BILL: I saw the dog [dag]

Courtesy of Dennis Preston, adapted from J. Schenkein (1978). *Studies in the organization of conversational interaction*. New York: Academic Press, pp. ix-xvi.

Appendix I
SAMPLE TRANSCRIPTION CONVENTIONS FOR THE L2 CLASSROOM

Transcription conventions for classroom discourse

General Layout

- 1 Leave generous margins at least at first to permit legible annotations as transcription gets refined.
- 2 Double space everything for the same reason.
- 3 Number every fifth line in the left-hand margin, *but* do so only in pencil until transcription is complete, unless you are using wordprocessing with automatic line numbering.
- 4 Identify transcripts at the top of each page with some economical reference number.
- 5 Number all pages in the top right corner.
- 6 Identify participants, date and location on a separate sheet (separate in case participants' identifies need to be kept confidential).
- 7 Decide whether to supply pseudonyms for participants' names, or to substitute numbers.
- 8 Enter participants' pseudonyms, where used, with gender, classroom layout, etc., also on a separate sheet (especially is using computer, since computer analysis must not include this page as data).
- 9 If using numbers, enter real name and associated numbers (with gender information) on a separate sheet.
- 10 On transcript pages, justify identifying material to the right, justify text to the left, as below.

Symbols to identify who is speaking

T	teacher
A	aide
M1	identified male learner, using numbers (M1, M2, etc.)
F1	identified female learner, using numbers (F1, F2, etc.)
Su	use such two-letter abbreviations for pseudonyms, where used (note: gender information may be lost by this method)
M	unidentified male learner
F	unidentified female learner
MV	male voice from, for example, an audio or videotape
FV	female voice, as above
LL	unidentified subgroup of class
<u>LL</u>	unidentified subgroup speaking in chorus
LLL	whole class
<u>LLL</u>	whole class speaking in chorus

Symbols for relationships between lines of transcript

{ M3
F7 use curly brackets to indicate simultaneous speech

[M
T
M use to indicate same unidentified male speaker

[F
T
F use to indicate same unidentified female speaker

-T use hyphen to indicate continuation of a turn without a pause, where overlapping speech intervenes.

Symbols to use in text

- [] use for commentary of any kind (e.g. to indicate point in discourse wehre T writes on blackboard)
- [=] use to introduce a gloss, or translation, of speech
- / / use for phonemic transcription instead of standard orthography, where pronunciation deviant. Use with gloss if meaning also obscured.
- () use for uncertain transcription
- (/ /) use for uncertain phonemic transcription
- ([]) use for uncertain gloss
- x incomprehensible item, probably one word only
- xx incomprehensible item of phrase length
- xxx incomprehensible item beyond phrase length
- x—x use optionally at early stages to indicate extent of incomprehensible item, as guide to future attempts to improve transcription
- use dots to indicate pauses, giving length in seconds in extreme cases, if potentially relevant to aims
- “ ” use to indicate anything read rather than spoken without direct text support

Further notes

- 1 Use indentation to indicate overlap of turns, otherwise start all turns systematically at extreme left of text space.
- 2 Use hyphen in text to indicate an incomplete word (for example, Come here, plea-)
- 3 Omit the full stop (period) at the end of a turn, to indicate incompleteness (for example, As I was going to)

OTHERWISE PUNCTUATE AS NORMALLY AS POSSIBLE, AS IF WRITING A PLAYSRIPT

- 4 Use 'uh' for hesitation fillers, or give phonemic transcription if meaning differences are potentially important.
- 5 Use underlining for emphasis, if using typewriter, or **bold** if wordprocessing (for example, Come **here!**)

GENERAL PRINCIPLE: THE LAW OF LEAST EFFORT

AVOID REDUNDANCY. Use only the conventions that are necessary for your particular purposes, to record the information you are sure you will need. If you are wordprocessing it will always be possible to update the transcript later (though admittedly this will be much more laborious if only typewriting facilities are available).

Allwright, D., & Bailey, K. M. (1991). *Focus on the language classroom: An introduction to classroom research for language researchers*. Cambridge: Cambridge University Press, pp. 222-223. Copyright © 1991 by Cambridge University Press. Reproduced with the permission of Cambridge University Press.

Appendix J
COMMONLY-USED FORMULÆ

Commonly-used Formulæ¹	
Statistic	Formula
Correlation	
Pearson product- moment	$\frac{\sum xy - (\sum x)(\sum y) / N}{\sqrt{[\sum x^2 - (\sum x)^2 / N][\sum y^2 - (\sum y)^2 / N]}}$
Spearman rho	$1 - \frac{6(\sum d^2)}{N(N^2 - 1)}$
Standard error of the mean (SEM)	$\frac{SD}{\sqrt{n-1}}$
Standard error of the difference (SED)	$\sqrt{(SEM_1)^2 + (SEM_2)^2}$

¹ This table uses some of the most commonly-used statistics. In most cases, a computer package (e.g., SPSS) will be used so there is no need to “know” the formula. We present these formulas to acquaint the reader with the kind of information that goes into the calculation of these common statistics.

t-test	$\frac{\bar{x}_1 - \bar{x}_2}{SED}$
paired t-test	$\sum \frac{\bar{x}_1 - \bar{x}_2}{s_D}$
ANOVA	$\frac{MSB}{MSW}$
MSB	$\frac{SSB}{K - 1}$
MSW	$\frac{SSW}{N - K}$
SST	$\frac{\sum x^2 - \frac{(\sum x)^2}{N}}$
Sum of squares between (SSB)	$\frac{[(\sum_{n_1} x_1)^2 + (\sum_{n_2} x_2)^2 + \dots + \sum_{n_k} (x_k)^2] - \frac{(\sum x)^2}{N}}$
Sum of squares within (SSW)	SST-SSB
χ^2	$\sum \frac{(f_o - f_e)^2}{f_e}$
z score	$\frac{x - \bar{x}}{sd}$
T score	$(10 z) + 50$

η^2 (for t-test)	$\frac{t^2}{t^2 + df}$
ω^2	$\frac{SSB - (K-1)MSW}{SST + MSW}$
Interrater reliability	$\frac{nr_{AB}^{**}}{1 + (n-1)r_{AB}}$
Cohen's d (effect size)	$\frac{\bar{x}^1 - \bar{x}^2}{S_w}$
Pooled standard deviation (S_w)	$\frac{(n_1-1)sd_1 + (n_2-1)sd_2}{(n_1-1) + (n_2-1)}$

* number of groups

** r_{AB} = correlation of two raters (if two) or average correlation if more than 2.

Glossary

Abstract	A brief summary of research that includes the research questions, the methods used (including participants) and the results.
Acceptability judgment	A judgment about the acceptability of a particular utterance (generally a sentence).
Action research	Generally refers to research carried out by practitioners in order to gain a better understanding of the dynamics of how second languages are learned and taught, together with some focus on improving the conditions and efficiency of learning and teaching.
Analysis of covariance (ANCOVA)	A type of analysis of variance that adjusts the measurement of the dependent variable to take other variables, such as a pretest score or an aptitude score, into account.
Analysis of variance	A parametric statistic that enables researchers to compare the performance between (generally) more than two groups.
Associational research	A research type that is concerned with co-occurrence and relationships between/ among variables.
Average	See mean.
Bell curve	See normal distribution.

Biodata	Basic information about a participant. The information gathered depends on the goal of a study. In general, age, amount, and type of prior L2 study, gender, first language of participant, and proficiency in L2s are collected and reported.
Case study	A detailed description of a single case, for example an individual learner or a class within a specific population and setting.
Chi square (χ^2)	A nonparametric statistic used with frequency data to test the relationship between variables.
CHILDES	A database of transcribed language acquisition data.
Classroom observation	An observation carried out in a classroom setting, often using a structured scheme or tally sheet for recording data.
Classroom research	Research conducted in second or foreign language classroom settings, often involving variables related to instruction.
Closed role play	Similar to discourse completion tasks, but in oral mode. Individuals are usually provided with a description of a situation and/or a character and asked to state what they would say in that particular situation. (See also Open role plays.)
Coding	Organizing data into a manageable, easily understandable, and analyzable base of information, and searching for and marking patterns in the data.
Coding system	A means of organizing data prior to analysis. Coding systems usually involve coding sheets, charts, techniques, schemes, and so on. Researchers develop their coding scheme based on their specific research questions.
Comparison group design	Compares performance following a treatment with two or more groups.

Computer-mediated communication (CMC)	Communicative exchanges between participants using a computer. Exchanges are recorded and information on performance such as keystrokes, erasures, and times can be documented.
Confirmability	Similar to the concept of replicability in quantitative research, confirmability in qualitative research involves making available full details of the data on which claims or interpretations are based so that other researchers can examine the data and confirm, modify or reject the first researcher's interpretations.
Consciousness-raising task	A task that is intended to facilitate learners' cognitive processes in terms of awareness of some language area or linguistic structure.
Consensus task	A task in which participants are presented with information to discuss, and if possible, come to agreement about something while utilizing that information.
Construct validity	The degree to which the research adequately captures the construct of interest.
Content validity	The extent to which a test or measurement device adequately measures the knowledge, skill or ability that it was designed to measure.
Control group design	A type of design that includes one group which does not receive the experimental treatment but participates in the testing sessions.
Control variable	A variable that is held constant across groups in order to eliminate the effect of that variable on the outcome of the study.
Convenience sample	A sample of the most available and/or accessible subjects in the population.
Corpus/Corpora	A collection of authentic data, often with detailed information about the context of collection and/or of use.
Correlation coefficient	A numerical value between +1 and -1 that indicates the strength of relationship between variables.
Correlational research	A type of research that involves data collection designed to determine the existence and strength of a relationship between two or more variables.

Counterbalancing	An experimental design in which the ordering of test items or tasks is different for different participants or groups of participants.
Covariate	A variable which is believed to influence the measurement of the dependent variable. Used in an analysis of covariance.
Credibility	A term used by qualitative researchers to ensure that the picture provided by the research is as full and complete as possible.
Criterion-related validity	The extent to which tests used in a study are comparable to other well-established tests of the construct in question.
Critical value	The value that is used as a confidence measure to determine whether a hypothesis can be substantiated or a null hypothesis can be rejected.
Cronbach's α	A means to determine internal consistency of a measure when only one administration of a measure exists. It is used when the number of possible answers is more than 2 and can be applied to ordinal data. (See also Kuder-Richardson 20 and split half procedure.)
Cyclical data analysis	A process where data collection is followed by some type of data analysis and hypothesis-formation, leading to subsequent and more focused rounds of data collection where hypotheses are tested and further refined, with the process continuing until a rich and full picture of the data is obtained.
Data	There are many forms of second language data. For example, data may be oral and recorded onto audio and/or videotapes; they may be written, in the form of essays, test scores, diaries, or check marks on observation schemes; they may appear in electronic format, such as responses to a computer-assisted accent modification program; or they may be visual, in the form of eye movements made while reading text at a computer or gestures made by a teacher in a classroom.

Data collection	The general process of accumulating information pertaining to a particular research question, problem or area.
Data elicitation	A subset of data collection, data elicitation refers to the process of directly eliciting information from individuals, for example, through an interview or a task.
Data sampling	Selecting and segmenting data, sometimes using only a portion of it in a procedure known as data reduction. Also known as data segmentation.
Data segmentation	See data sampling.
Debriefing	Providing information after a study or data collection period. For example, participants may be informed about research findings, questions, or the content of observations.
Degrees of freedom	The number of scores that can vary if others are given.
Delayed posttest	In a pretest/posttest design, delayed posttests are tests given after the first posttest (for example, 1 month or 1 year later) to measure the long-term retention of a skill or knowledge.
Dependability	Similar to consistency and reliability in quantitative research, the extent to which people not involved in the study would make the same observations and draw the same conclusions when following the same research steps.
Dependent variable	The variable that is measured to determine the effects of the independent variable.
Diary research	An individual's perspective on their own language learning or teaching experience, in the form of entries to a personal journal. Analyses usually focus on patterns and salient events.
Directional hypothesis	A prediction that specifies the relationship between variables. This is generally stated in the form of X will be greater than Y. (See one-way hypothesis.)

Discourse completion test (DCT)	DCT's are a means of gathering contextualized data. Generally, a situation is provided and then the respondent is asked what she or he would say in that particular situation. There is often a follow-up response (such as "I'm sorry that you can't come) so that the individual knows the type of response that is expected (for example, a refusal).
Distribution	A way of showing the frequency with which scores occur in a data set.
Duncan's multiple range test	A post-hoc test used to compare means following an analysis of variance.
Dyad	Two participants working together.
Effect of instruction research	A kind of classroom research that focuses on the role and learning outcomes of (different types of) second language instruction.
Effect size	A measure that can be used to determine the magnitude of an observed relationship or effect.
Elicited imitation	A procedure for collecting data where a participant is presented with a sentence, clause or word and is asked to repeat (imitate) it.
Elicited narrative	Narratives that are gathered through specific prompts (e.g., What did you do yesterday? Or, Tell me about a typical day for you.)
Emic	An insider's understanding of his or her own culture.
Empirical research	Research that is based on data.
Eta ²	A correlation coefficient that expresses the strength of association and can be used following a t-test. It is expressed as η^2 .
Ethics review board	See Institutional Review Board.
Ethnography	Research that is carried out from the the participants' point of view, using categories relevant to a particular group and cultural system. It aims to describe and interpret the cultural behavior, including communicative patterns, of a group.

Etic	An outsider's understanding of a culture or group that is not their own.
Experimental research	Research in which there is manipulation of (at least) one independent variable to determine the effect(s) on one (or more) dependent variables. Groups are determined on the basis of random assignment.
External validity	Refers to extent to which the results of a study are relevant to a wider population.
Face validity	Refers to the familiarity of an instrument and the ease with which the validity of the content is recognized.
Factor analysis	A means of determining common factors that underlie measures that test different (possibly related) variables. It allows the researcher to take a larger number of variables and reduce them to a smaller number of factors.
Factorial design	A design type that involves more than one independent variable. The goal is to determine the effects of each individually and in interaction on the dependent variable.
Fisher's exact test	A variant of the chi square test (χ^2) used for 2×2 contingency tables. Fisher's exact test can be used with low frequency counts in cells.
Focus groups	Related to interviews, these involve several participants in a group discussion, often with a facilitator whose goal it is to keep the group discussion targeted on specific topics. A stimulus is generally used for discussion, such as a videotape or previously elicited data.
Friedman test	A non-parametric test used to compare three or more matched groups. This test is the nonparametric equivalent with repeated measures ANOVA.
Generalizability	The extent to which the results of a study can be extended to a greater population.

Grounded theory	Theory based on, or grounded in, data that has been systematically gathered and analyzed. Researchers often attempt to avoid placing preconceived notions on the data, and aim to examine data from multiple vantage points to help them arrive at a complete picture of the phenomena under investigation.
Halo effect	Participants provide information that they believe a researcher wants or expects.
Hawthorne effect	The presence of observers may result in changed behavior due to the fact that those being observed feel positive about being included in a study.
Human subjects committee	See Institutional Review Board.
Hypothesis	A statement of what one believes the outcomes of a study will be. A research hypothesis predicts what the relationship will be between or among variables.
Independent variable	A variable that is believed to affect the dependent variable.
Inductive data analysis	The general goal is for research findings to emerge from the frequent, dominant or significant themes within raw data. Involves multiple examinations and interpretations of the data in the light of the research objectives.
Inferential statistics	A type of statistic that determines the likely generalizability from a sample(s) to the general population.
Information-exchange tasks	A task in which two (or more) individuals must exchange information using the linguistic resources available to them in order to complete the activity. A spot the differences task, where each participant has a uniquely held picture and they must share information to complete the task is a type of information-gap task, also known as a jigsaw task.

Information-gap tasks	In an information-gap task, one individual usually has a gap in his/her information. For example, a picture drawing activity, where one person describes and another person draws, is a type of information gap task.
Informed consent	Voluntary agreement to participate in a study about which the potential subject has enough information and understands enough to make an informed decision.
Institutional Review Board	A committee established to review research involving human subjects to ensure it is in compliance with ethical guidelines laid down by government and funding agencies. (This term is often used interchangeably with Human Subjects Committee and Ethics Review Board.)
Instrument reliability	Refers to the consistency of a particular instrument over time.
Intact class	A treatment group that is made up of all individuals in a given class.
Interaction effect	Associated with a factorial design. Combined effect of two independent variables (see main effect).
Internal validity	The extent to which the results of a study are a function of the factor that is intended by the researcher.
Interrater reliability	Consistency between two or more raters.
Interval scale	A scale in which there is an ordering of variables and in which there is an equal interval between variables.
Intervening variable	A variable that is not controlled for but can have an effect on the relationship between the independent and dependent variables.
Interview	Comparable to a questionnaire, but in oral mode. Interviews are often associated with survey-based research. Information is often gathered by means of open-ended questions and answers. Interviews can be based around a stimulus, for example a completed questionnaire, or a videotape of a lesson.

Intrarater reliability	A single rater's consistency at two or more points in time.
Introspective methods	A set of data elicitation techniques that encourage learners to communicate about their internal processing and/or perspectives about language learning experiences.
Investigator triangulation	Using multiple observers or interviewers in the same investigation.
Kendall Tau	A correlation analysis used for ordinal data or with interval data when converted to ranks. This analysis is used when there are many ties in the rankings. (See Spearman rho.)
Kruskal-Wallis	A non-parametric test used to compare two or more independent groups.
Kuder-Richardson 20	A means to determine the internal consistency of a measure. It is used when one administration of a measure exists. This is done when the instrument is not split into two parts. (See also split half and Cronbach's α .)
Laboratory research	Generally taken to refer to experimental research, where variables can be manipulated and the setting is controlled.
Linear regression	A means of predicting the score on one variable from another score on another variable.
Magnitude estimation	A procedure whereby participants are asked to rank a stimulus by stating how much better or worse the stimulus is from the previous one.
Main effect	The effect of one independent variable (see interaction effect).
Mann Whitney U	A non-parametric test used with ordinal or interval data. It is used to compare two groups and is similar to the Wilcoxon Rank sums test.
Mean	A measure of central tendency, the mean is a value obtained by summing all the scores in a score distribution and dividing the sum by the number of scores in the distribution.

Measures of central tendency	A way of providing quantitative information about the typical behavior of individuals with respect to a particular phenomenon, usually given through means, modes and medians.
Measures of dispersion	A method for determining the amount of spread in a set of scores. It is a way of determining variability.
Measures of frequency	Indicate how often a particular behavior or phenomenon occurs.
Median	A measure of central tendency, the median is a value that represents the mid-point of all the scores. Half of the scores are above the median and half below it.
Meta-analysis	A statistical tool used in research synthesis to convert the findings of individual studies to comparable values in order to estimate the overall observed finding about a given treatment or condition across studies.
Methodological triangulation	Using different independent measures or methods to investigate a particular phenomenon.
Mode	A measure of central tendency, the mode represents the most frequent score obtained in a score distribution.
Moderator variable	A variable that may interact with other variables resulting in an effect on the relationship between the independent and dependent variables..
Moving window	A technique, generally carried out on a computer, whereby words are presented visually or aurally one by one. This technique may be used to measure reaction times (i.e., how long it takes a participant to press a button to have the next word presented) and thus, indirectly, ease of comprehension.
Multiple methods research	See split methods research.
Multiple regression	A means of predicting a score from the scores of two or more variables.
Multivariate analysis of variance (MANOVA)	A type of analysis of variance with more than one dependent variable.

Naturalistic data	Data that come from naturally occurring situations and which are not experimentally manipulated.
Naturalistic setting	A research context that involves no manipulation of variables. Data are collected through observations of settings in which data occur without specific intervention or control.
Nominal scale	Used for attributes or categories. A nominal scale is used to place attributes into two or more categories (e.g., gender).
Non-directional hypothesis	A prediction that a relationship between variables exists without specifying the precise nature of the direction. (See two-way hypothesis.)
Non-parametric statistics	A type of inferential statistics that are generally used with nominal or ordinal data or when the assumptions necessary for parametric statistics cannot be met.
Normal distribution	A theoretical distribution of scores. In a normal distribution, scores cluster around the mid-point.
Observations	Researchers systematically observe different aspects of a setting in which they are immersed, including, for example the interactions, relationships, actions, and events in which learners engage. The aim is to provide careful descriptions of learners' activities without unduly influencing the events in which the learners are engaged.
Ω^2	A measure used to determine the strength of association when all groups have an equal n size.
One-shot design	A design type that uses one treatment and one measurement afterwards; there is no measurement before the treatment (e.g., a pretest) and there is no control group.
One-way hypothesis	A prediction that specifies the relationship between variables. This is generally stated in the form of X will be greater than Y. (See directional hypothesis.)
On-line task	See think-aloud.

Open role-play	Individuals are provided with a description of a situation and/or a character and each individual is asked to play out the part of one of the characters. In open role-plays, limits are not provided as to the length of the exchange. (See also closed role-play.)
Operationalize	To provide a precise, concrete definition of a variable in such a way that it can be measured.
Ordinal scale	A scale in which there is an ordering of variables. There is no implication that there is an equal interval between variables.
Outlier	A score that is different from the other scores in a set. It may be considerably larger or smaller than all the other scores.
Paired t-test	A type of t-test used when the comparison is between matched samples (e.g., pretest-posttest). It is also known as a matched t-test.
Parameters	A way of describing certain characteristics of a population numerically.
Parametric statistics	Inferential statistics that use sets of assumptions about the dependent variable. Generally used with interval data.
Participant	An individual whose behavior is being measured or investigated.
Participant mortality	The drop out rate for a study. It is also referred to as subject mortality. Participants drop out for many reasons including scheduling conflicts (there is often a high rate of no-shows for delayed post tests).
Pearson product-moment correlation	A measure of correlation used with interval data.
Pilot study	A small-scale trial of the proposed procedures, materials, and methods. It may also include a trial of the coding sheets and analytic categories.
Population	All instances of individuals (or situations) that share certain characteristics.

Post-hoc analysis	A follow-up statistical analysis performed after a comparison of more than two groups (e.g., Analysis of variance) shows a significant difference. It is a way of pinpointing where, for example between which groups or tests, the significant difference lies.
Posttest	A test to determine knowledge after treatment.
Posttest only design	Uses one treatment and one measurement afterwards. Like a one-shot design, there is no measurement before the treatment (e.g., a pretest); however, there is usually a control group.
Practitioner research	See Action research.
Predictive validity	Refers to the use that one wants to make of a measure. If there is predictive validity, the measure can predict performance on some other measure.
Pretest	A test to determine knowledge before treatment.
Pretest/posttest design	Compares performance before treatment with performance following treatment.
Probability	An estimation of the likelihood of something occurring due to chance.
Purpose sample	A sample selected in order to elicit a particular type of data. The sample may or may not be representative of the population at large.
Qualitative research	Research in which the focus is on naturally occurring phenomena and data are primarily recorded in non-numerical form.
Quantification	The use of numbers and sometimes statistics to show patterns of occurrence.
Quantitative research	Research in which variables are manipulated to test hypotheses and in which there is usually quantification of data and numerical analyses.
Quasi-experimental research	A type of experimental research but without random assignment of individuals.

Questionnaire	A (usually written) survey often used in a large-scale study to gather information. Can utilize open-ended questions and/or questions followed by a selection from a set of pre-determined answers.
Random sample	A sample that has been selected in such a way that each member of a population has an equal chance of being selected. (See simple and stratified random sampling.)
Range	A measure of dispersion, range indicates the distance between the highest and lowest score. It measures the spread of a set of scores.
Ratio scale	An interval scale that displays information about frequencies in relation to each other
Reaction time	The time between a stimulus and a learner's response. Reaction time experiments are usually computer-based and can also be used to investigate processing.
Regression line	A line that can be drawn through scores on a scatterplot. It is the line of best fit, that is, the one which allows for the clustering of scores on the line.
Reliability	The degree to which there is consistency in results.
Repeated measures design	Multiple measurements from each participant.
Replication	Conducting a research study again, in a way that is either identical to the original procedure or with small changes (e.g., different participants), to test the original findings.
Representativeness	The extent to which an individual who could be selected for a study has the same chance of being selected as any other individual in the population
Research	A systematic process of collecting and analyzing information that will investigate a research problem or question, or help researchers obtain a more complete understanding of a situation.

Research protocol	A detailed set of guidelines for how research will proceed.
Research question	A question that will be addressed/investigated in a study.
Research report	A formal report in which the findings from research are presented.
Sampling	The way participants or data for a study are selected.
SAS	A statistical package for data analysis.
Scheffé	A post-hoc test used to compare means following an analysis of variance.
Semi-structured interview	An interview in which researchers use written lists of questions as a guide, but can digress and probe for more information.
Sentence matching	A procedure (generally computer-based) whereby participants are asked if two sentences (usually appearing consecutively) are identical or not. This procedure is often used to determine grammaticality.
Simple random sample	Refers to a sample that has been selected in such a way that each member of a population has an equal chance of being selected
Small group	Usually three or four participants; however, depending on the context, five may also be considered a small group.
Spearman rho	A correlation analysis used for ordinal data or with interval data when converted to ranks. (See Kendall Tau.)
Split half procedure	A means of determining internal consistency of a measure by obtaining a correlation coefficient by comparing the performance on half of a test with performance on the other half. (See also Kuder-Richardson 20 and Cronbach's α .)

Split methods research	Research where authors present and discuss both quantitative and qualitative data in the same report, or use methods associated with both types of research in collecting data or conducting studies. (Sometimes known as multiple methods research or mixed methods research.)
SPSS	A statistical package for data analysis.
Standard deviation	A measure of dispersion, it is a numerical value that indicates how scores are spread around the mean.
Standard error of the difference between sample means	Difference between sample means.
Standard error of the mean	Standard deviation of sample means.
Standard scores	A converted raw score that shows how far an actual score is from the mean.
Standardized interview	See structured interview.
Statistic	A way of describing a sample numerically.
Stimulated recall	An introspective technique for gathering data that can yield insights into a learner's thought processes during language learning experiences. Learners are asked to introspect while viewing or hearing a stimulus to prompt their recollections.
Stratified random sample	Random sampling based on categories
Strength of association	A way of determining how much variation in the data can be accounted for by the independent variable
Structured interview	Researchers ask similar sets of questions of all respondents. Structured interviews resemble verbal questionnaires. (Also known as standardized interviews.)
Suppliance in obligatory contexts (SOC) measures	Whether or not a feature is produced when it is required. Usually, the number of times a particular feature (e.g., the past tense) is produced is divided by the number of times that it is required.

Survey	A means of gathering information about a particular topic, for example, attitudes or opinions about a school program. A questionnaire is a type of survey.
SYSTAT	A statistical package for data analysis.
Systematic sample	A sample that has been determined by the selection of every <i>n</i> th individual or instance / occurrence for sampling data.
T score	A standard score based on a <i>z</i> score (multiply <i>z</i> score by 10 and add 50).
Teacher-initiated research	See Action research.
Theoretical triangulation	Using multiple perspectives to analyze the same set of data.
Thick description	Using multiple perspectives to explain the insights gleaned from a study, and taking into account the participants interpretations of their actions and speech. It also involves the presentation of representative examples and patterns in data, along with interpretive commentary.
Think-aloud	A type of verbal reporting in which individuals are asked what is going through their mind as they are solving a problem or performing a task. (See also on-line tasks.)
Time series design	A design type that involves repeated observations over a set period of time where the participants serve as their own control.
Transcription conventions	Notations used to facilitate the representation of oral data in a written format. While there are no generally agreed-upon conventions common to all studies, researchers may recognize certain symbols; for instance, the use of dots to convey pauses or silence is relatively common.

Transcription machines	Designed to facilitate transcription of oral data, these machines are often controlled by a foot pedal so that both hands are free for typing. As well as rewinding by a set number of seconds, the controls can be used to adjust the rate of the speech, to make it easier to distinguish individual voices.
Transferability	How far qualitative research findings are transferable from one context to another. The extent to which findings may be transferred usually depends on the similarity of the context.
Triangulation	Triangulation involves using multiple research techniques and multiple sources of data in order to explore the issues from all feasible perspectives. Using the technique of triangulation can aid in credibility, transferability, confirmability, and dependability in qualitative research.
Truth-value judgment	These judgments generally involve contextualized information and individuals are asked if a particular follow-up sentence is true or not based on prior contextualization.
t-test	A parametric statistic that is used to determine if the means of two groups are significantly different from one another. (See also paired t-test.)
T-unit	Usually defined as one main clause and any attached dependent clauses.
Tukey	A post-hoc test used to compare means following an analysis of variance.
Two-way hypothesis	A prediction that a relationship between variables exists without specifying the precise nature of the direction. (See non-directional hypothesis.)
Type I error	A null hypothesis is rejected when it should not have been rejected.
Type II error	A null hypothesis is accepted when it should not have been accepted.

Type-token ratio	A measure of lexical diversity which involves dividing the number of types by the number of tokens. For example types can refer to the different words that are used in one data set, and tokens can refer to the number of repetitions of those words.
Uptake sheets	Learners' reports about their learning, illustrating what they take up from the language learning opportunities they have through instruction.
Validity	The extent one can make correct generalizations based on the results from a particular measure.
VARBRUL	A statistical package for data analysis often used in sociolinguistic research.
Variable	A characteristic that differs from group to group or person to person (e.g., native language, handedness).
Verbal reporting	A type of introspection that consists of gathering information by asking individuals to say what is going through their minds as they are solving a problem or doing a task.
Wilcoxon Rank Sums	A non-parametric test used with ordinal or interval data. It is used to compare two groups and is similar to the Mann Whitney U test.
Within-group design	See repeated measures design.
z scores	A standard score that provides information about the distance of a score from the mean in terms of standard deviations.

References

- Adair, G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2), 334–345.
- Adams, R. (2003). L2 output, reformulation, and noticing: Implications for IL development. *Language Teaching Research*, 7, 347–376.
- Akiyama, Y. (2002). Japanese adult learners' development of the locality condition on English reflexives. *Studies in Second Language Acquisition*, 24, 27–54.
- Allen, J. P. B., Fröhlich, M., & Spada, N. (1984). The communicative orientation of language teaching: An observation scheme. In J. Handscombe, R. A. Orem, & B. Taylor (Eds.), *On TESOL '83: The question of control* (pp. 231–252). Washington, DC: TESOL.
- Allwright, D. (1984a). Why don't learners learn what teachers teach? The interaction hypothesis. In D. Singleton & D. Little (Eds.), *Language learning in formal and informal contexts* (pp. 3–18). Dublin: IRAAL.
- Allwright, R. (1984b). The importance of interaction in classroom language learning. *Applied Linguistics*, 5, 158–171.
- Allwright, D. (1987). Classroom observation: Problems and possibilities. In B. K. Das (Ed.), *Patterns of classroom interaction in Southeast Asia* (pp. 88–102). Singapore: SEAMEO Regional Language Center.
- Allwright, D., & Bailey, K. M. (1991). *Focus on the language classroom: An introduction to classroom research for language researchers*. Cambridge, UK: Cambridge University Press.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- The American Heritage Dictionary of the English Language, Fourth Edition*. (2000). Boston: Houghton Mifflin.
- Atkinson, D., & Ramanathan, V. (1995). Cultures of writing: An ethnographic comparison of L1 and L2 university writing programs. *TESOL Quarterly*, 29, 539–568.
- Bailey, K. M. (1983). Competitiveness and anxiety in adult second language learning: Looking at and through the diary studies. In H. W. Seliger & M. H. Long

- (Eds.), *Classroom oriented research in second language acquisition* (pp. 67–103). Rowley, MA: Newbury House.
- Bailey, K. M. (1990). The use of diary studies in teacher education programmes. In J. Richards & D. Nunan (Eds.), *Second language teacher education* (pp. 215–226). Cambridge, UK: Cambridge University Press.
- Bailey, K. M., Bergthold, B., Braunstein, B., Fleischman, N. J., Holbrook, M. P., & Zambo, L. J. (1996). The language learner's autobiography: Examining the "apprenticeship of observation." In D. Freeman & J. C. Richards (Eds.), *Teacher learning in language teaching* (pp. 11–29). Cambridge, UK: Cambridge University Press.
- Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *The Modern Language Journal*, 87, 546–561.
- Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72, 32–68.
- Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, 26(2), 390–395.
- Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Oxford: Blackwell.
- Bardovi-Harlig, K., & Dörnyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, 32, 233–262.
- Bardovi-Harlig, K., & Hartford, B. S. (1996). Input in an institutional setting. *Studies in Second Language Acquisition*, 18, 171–188.
- Bates, E., & MacWhinney, B. (1982). Functionalist approach to grammar. In E. Warmer & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 173–218). New York: Cambridge University Press.
- Baumrind, D. (1990). Doing good well. In C. B. Fisher & W. W. Tryon (Eds.), *Ethical issues in applied developmental psychology* (pp. 17–28). Norwood, NJ: Ablex.
- Beck, M. L. (1998). German and the local impairment hypothesis. *Studies in Second Language Acquisition*, 20, 311–348.
- Beebe, L., & Takahashi, T. (1989). Do you have a bag? Social status and patterned variation in second language acquisition. In S. Gass, C. Madden, D. Preston, & L. Selinker (Eds.), *Variation in second language acquisition: Discourse and pragmatics* (pp. 103–125). Clevedon, UK: Multilingual Matters.
- Beebe, L., Takahashi, T., & Uliss-Weltz, R. (1990). Pragmatic transfer in ESL refusals. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 55–73). New York: Newbury House.
- Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*, 8(3), 215–241.
- Bernard, H. R. (1995). *Research methods in anthropology: Qualitative and quantitative approaches*. Walnut Creek, CA: AltaMira.
- Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence*. New York: Springer.

- Bley-Vroman, R., Felix, S., & Ioup, G. (1988). The accessibility of Universal Grammar in adult language learning. *Second Language Research*, 4, 1–32.
- Bley-Vroman, R., & Joo, H-R. (2001). The acquisition and interpretation of English locative constructions by native speakers of Korean. *Studies in Second Language Acquisition*, 23, 207–219.
- Bley-Vroman, R., & Masterson, D. (1989). Reaction time as a supplement to grammaticality judgments in the investigation of second language learners' competence. *University of Hawai'i Working Papers in ESL*, 8, 207–245.
- Breen, M. P., Hird, B., Milton, M., Oliver, R., & Thwaite, A. (2001). Making sense of language teaching: Teachers' principles and classroom practices. *Applied Linguistics*, 22(4), 470–501.
- British Association for Applied Linguistics. (1994). *Recommendations on good practice in applied linguistics*. Retrieved January 24, 2004, from <http://www.baal.org.uk/goodprac.htm#6>
- Brown, J. A. C. (1954). *The social psychology of industry*. Middlesex, UK: Penguin.
- Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.
- Brown, J. D. (2003). Research methods for applied linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp 476–500). Oxford, UK: Blackwell.
- Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Butler, C. (1985). *Statistics in linguistics*. Oxford, UK: Basil Blackwell.
- Carmines, E., & Zeller, R. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.
- Carson, J., & Nelson, G. L. (1996). Chinese students' perception of ESL peer response group interaction. *Journal of Second Language Writing*, 5, 1–19.
- Chalmers, D. (2003, July). *Ethical and policy issues in research involving human participants commissioned papers and staff analysis: Research ethics in Australia*. Retrieved January 24, 2004, from <http://onlineethics.org/reseth/nbac/hchalmers.html#system>
- Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning*. Cambridge, UK: Cambridge University Press.
- Chaudron, C. (2000). Contrasting approaches to classroom research: Qualitative and quantitative analysis of language and learning. *Second Language Studies*, 19(1), 1–56. Available online at http://www.hawaii.edu/sls/uhwpsel/19_1/chaudron.pdf
- Chaudron, C., Crookes, G., & Long, M. H. (1988). *Reliability and validity in second language classroom research*. Honolulu: Center for Second Language Classroom Research, Social Science Research Institute, University of Hawaii.
- Chomsky, N. (1997). *The minimalist program*. Cambridge, MA: MIT Press.

- Cohen, A. (1998). *Strategies in learning and using a second language*. London: Longman.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen J. D., MacWhinney, B., Flatt, M., & Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, 25(2), 257–271.
- Crookes, G. (1993). Action research for second language instructors: Going beyond instructor research. *Applied Linguistics*, 14(2), 130–144.
- Davis, K. A. (1995). Qualitative theory and methods in applied linguistics research. *TESOL Quarterly*, 29, 427–453.
- de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19, 249–276.
- De Guerrero, M., & Villamil, O. (2000). Activating the ZPD: Mutual scaffolding in L2 peer revision. *The Modern Language Journal*, 84(1), 56–68.
- Denzin, N., & Lincoln, Y. (Eds.). (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.
- Department of Health and Human Services (DHHS). (2003). *Protection of human subjects 45, Code of Federal Regulations (CFR) 46, Subparts A and D*. Washington, DC: U.S. Government Printing Office.
- Derwing, T., Munro, M., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48, 393–410.
- Dörnyei, Z. (2003). *Questionnaires in second language research: Constructing, administering, and processing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Dörnyei, Z., & Clément, R. (2001). Motivational characteristics of learning different target languages: Results of a nationwide survey. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 399–426). Honolulu: Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.
- Duff, P. (2002). The discursive co-construction of knowledge, identity and difference: An ethnography of communication in the high school mainstream. *Applied Linguistics*, 23, 289–322.
- Duff, P., & Early, M. (1996). Problematics of classroom research across sociopolitical contexts. In J. Schachter & S. Gass (Eds.), *Second language classroom research: Issues and opportunities* (pp. 1–30). Mahwah, NJ: Lawrence Erlbaum Associates.
- Duffield, N., Prévost, P., & White, L. (1997). A psycholinguistic investigation of clitic placement in second language acquisition. In E. Hughes, M. Hughes, & A. Greenhill (Eds.), *Proceedings of the 21st Boston University Conference on Language Development* (pp. 148–159). Somerville, MA: Cascadilla Press.

- Duffield, N., & White, L. (1999). Assessing L2 knowledge of Spanish clitic placement: Converging methodologies. *Second Language Research*, 15, 133–160.
- Egi, T. (2003, October). *Learners' perceptions about recasts and L2 learning*. Paper presented at the Second Language Research Forum, Tucson, AZ.
- Ellis, N. (2000). Editor's statement. *Language Learning*, 50, xi–xii.
- Ellis, R. (1984). *Second language classroom development*. Oxford, UK: Pergamon.
- Ellis, R. (1990). *Instructed second language acquisition*. Cambridge, MA: Blackwell.
- Ellis, R. (1994). *The study of second language acquisition*. Oxford, UK: Oxford University Press.
- Ellis, R. (2000). Does form-focused instruction affect the acquisition of implicit knowledge? A review of the research. *Studies in Second Language Acquisition*, 24, 223–236.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, UK: Oxford University Press.
- Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, 51(2), 281–318.
- Ellis, R., & He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, 21, 285–301.
- Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26, 59–84.
- EndNote (Version 7.0) [Computer software]. Carlsbad, CA: Thomson ISI ResearchSoft.
- Eubank, L. (1993). Sentence matching and processing in L2 development. *Second Language Research*, 9, 253–280.
- Fanselow, J. F. (1977). Beyond "Rashomon"—conceptualizing and describing the teaching act. *TESOL Quarterly*, 11, 17–39.
- Fiksdal, S. (1990). *The right time and place: A microanalysis of cross-cultural gatekeeping interviews*. Norwood, NJ: Ablex.
- Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics*, 19, 1–23.
- Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, 25, 605–628.
- Fraenkel, J., & Wallen, N. (2003). *How to design and evaluate research in education* (5th ed.). New York: McGraw-Hill.
- Francis, W., & Kucera, H. (1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin.
- Fukkink, R., Blok, H., & de Glopper, K. (2001). Deriving word meaning from written context: A multicomponential skill. *Language Learning*, 51, 477–496.
- Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly*, 14(1), 53–60.

- Gass, S. (1980). An investigation of syntactic transfer in adult L2 learners. In R. Scarcella & S. Krashen (Eds.), *Research in second language acquisition* (pp. 132–141). Rowley, MA: Newbury House.
- Gass, S. (1994). The reliability of second-language grammaticality judgments. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 303–322). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S. (2001). Sentence matching: A reexamination. *Second Language Research*, 17, 421–441.
- Gass, S., & Alvarez-Torres, M. (2005). Attention when?: An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition*, 27(1), 1–31.
- Gass, S., & Houck, N. (1999). *Interlanguage refusals: A cross-cultural study of Japanese-English*. Berlin: Mouton de Gruyter.
- Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S., Mackey, A., Alvarez-Torres, M., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49, 549–581.
- Gass, S., Mackey, A., & Ross-Feldman, L. (2003, October). *The role of setting in classroom and laboratory interaction: From claims to data*. Paper presented at the Second Language Research Forum, Tucson, AZ.
- Gass, S., & Selinker, L. (2001). *Second language acquisition: An introductory course*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S., Svetics, I., & Lemelin, S. (2003). Differential effects of attention. *Language Learning*, 53, 495–543.
- Gass, S., & Varonis, E. (1994). Input, interaction and second language production. *Studies in Second Language Acquisition*, 16, 283–302.
- Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.
- Georgetown University. (n.d.). *Institutional review board application forms*. Retrieved August 3, 2004, from <http://www.georgetown.edu/grad/IRB/irb-forms.html>
- Glew, M. (1998). *The acquisition of reflexive pronouns among adult learners of English*. Unpublished doctoral dissertation, Michigan State University, East Lansing.
- Goldschmidt, M. (1996). From the addressee's perspective: Imposition in favor-asking. In S. Gass & J. Neu (Eds.), *Speech acts across cultures* (pp. 241–256). Berlin: Mouton de Gruyter.
- Goldstein, T. (1997). *Two languages at work: Bilingual life on the production floor*. New York: Mouton de Gruyter.
- Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: John Benjamins.

- Grotjahn, R. (1987). On the methodological basis of introspective methods. In C. Færch & G. Kasper (Eds.), *Introspection in second language research* (pp. 54–81). Clevedon, UK: Multilingual Matters.
- Gwet, K. (2001). *Handbook of inter-rater reliability*. Gaithersburg, MD: Stataxis.
- Hall, A., & Rist, R. (1999). Integrating multiple qualitative research methods (or avoiding the precariousness of a one-legged stool). *Psychology and Marketing*, 16(4), 291–304.
- Harrington, M. (1987). Processing transfer: Language-specific strategies as a source of interlanguage variation. *Applied Psycholinguistics*, 8, 351–378.
- Harvard University. (2000, July). *The intelligent scholar's guide to the use of human subjects in research*. Retrieved April 15, 2002, from <http://www.fas.harvard.edu/~research/ISGnew.html>
- Hawkins, R., & Chan, C. Y. (1997). The partial availability of Universal Grammar in second language acquisition: The “failed functional features hypothesis.” *Second Language Research*, 13, 187–226.
- Heath, A. W. (1997, March). The proposal in qualitative research. *The qualitative report*, 3(1). Retrieved December 2, 2003, from <http://www.nova.edu/ssss/QR/QR3-1/heath.html>
- Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge, UK: Cambridge University Press.
- Helms-Park, R. (2001). Evidence of lexical transfer in learner syntax: The acquisition of English causatives by speakers of Hindi-Urdu and Vietnamese. *Studies in Second Language Acquisition*, 23, 71–102.
- Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality*, 5, 41–49.
- Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, 19, 131–143.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Report No. 3). Urbana, IL: National Council of Teachers of English.
- Inagaki, S. (2001). Motion verbs with goal PPs in the L2 acquisition of English and Japanese. *Studies in Second Language Acquisition*, 23, 153–170.
- Indiana University Office of the Vice President for Research. (1999–2003). *Bloomington Campus Committee for the Protection of Human Subjects: Students as subjects*. Retrieved January 19, 2004, from <http://www.indiana.edu/%7Eeresrisk/stusub.html>
- Ionin, T., & Wexler, K. (2002). Why is “is” easier than “s”? Acquisition of tense/agreement morphology by child second language learners of English. *Second Language Research*, 18, 95–136.
- Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24, 617–637.
- Jick, T. (1984). Mixing qualitative and quantitative methods: Triangulation in action. In J. Van Maanen (Ed.), *Qualitative methodology* (pp. 135–148). Beverly Hills: Sage.

- Johnson, D. M. (1992). *Approaches to research in second language learning*. New York: Longman.
- Johnson, D. M. (1993). Classroom-oriented research in second-language learning. In A. O. Hadley (Ed.), *Research in language learning: Principles, processes, and prospects* (pp. 1–23). Lincolnwood, IL: National Textbook Company.
- Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of ESL. *Cognitive Psychology*, 21, 60–99.
- Juffs, A. (2001). Discussion: Verb classes, event structure, and second language learners' knowledge of semantics–syntax correspondences. *Studies in Second Language Acquisition*, 23, 305–313.
- Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing: Subject and object asymmetries in wh-extraction. *Studies in Second Language Acquisition*, 17, 483–516.
- Just, M. A., Carpenter, P. A., & Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 3, 228–238.
- Kasper, G., & Rose, K. (2002). Pragmatic development in a second language. *Language Learning*, 52(Suppl. 1).
- Kennedy, E. (1988). *The oral interaction of native speakers and non-native speakers in a multicultural preschool: A comparison between freeplay and contrived NS/NNS dyads*. Unpublished master's thesis, University of British Columbia, Vancouver.
- King, K. A. (2000). Language ideologies and heritage language education. *International Journal of Bilingual Education and Bilingualism*, 3(2), 167–184.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Lakshmanan, U. (1989). *Accessibility to Universal Grammar in child second language acquisition*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.
- Lakshmanan, U., & Teranishi, K. (1994). Preferences versus grammaticality judgments: Some methodological issues concerning the governing category parameter in second language acquisition. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 185–206). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Leow, R. (1998). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) fine-grained analysis of attention. *Applied Psycholinguistics*, 19, 133–159.
- Leow, R. (2000). A study of the role of awareness in foreign language behavior: Aware vs. unaware learners. *Studies in Second Language Acquisition*, 22, 557–584.
- Leow, R., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, 26, 35–57.
- Lieven, E. V. M. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway & B. J. Richards (Eds.), *Input and inter-*

- action in language acquisition* (pp. 56–73). Cambridge, UK: Cambridge University Press.
- Lightbown, P. M., & Spada, N. (1990). Focus-on-form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition*, 12(4), 429–448.
- Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL Quarterly*, 28(3), 563–579.
- Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.
- Lotto, L., & de Groot, A. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, 48, 1–29.
- Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge, UK: Cambridge University Press.
- Lyster, R. (1998a). Form in immersion classroom discourse: In or out of focus? *Canadian Journal of Applied Linguistics*, 1, 53–82.
- Lyster, R. (1998b). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition*, 20, 51–81.
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1), 37–66.
- Macaro, E. (2001). *Learning strategies in foreign and second language classrooms*. London: Continuum.
- Macias, J. (1987). The hidden curriculum of Papago teachers: American Indian strategies for mitigating cultural discontinuity in early schooling. In G. Spindler & L. Spindler (Eds.), *Interpretive ethnography of education: At home and abroad* (pp. 363–380). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, 21, 557–587.
- Mackey, A. (2000, August). *Feedback, noticing and second language development: An empirical study of L2 classroom interaction*. Paper presented at BAAL 2000, Cambridge, UK.
- Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive implicit negative feedback? *Studies in Second Language Acquisition*, 22, 471–497.
- Mackey, A., McDonough, K., Fujii, A., & Tatsumi, T. (2001). Learners' reports about the L2 classroom. *International Review of Applied Linguistics*, 39, 285–307.
- Mackey, A., & Oliver, R. (2002). Interactional feedback and children's L2 development. *System*, 30, 459–477.
- Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal*, 82(3), 338–356.

- MacWhinney, B. (1999). The CHILDES system. In W. C. Ritchie & T. Bhatia (Eds.), *Handbook of child language acquisition* (pp. 457–494). New York: Academic Press.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Markee, N. (2000). *Conversation analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mason, J. (1996). *Qualitative researching*. London: Sage.
- Mayo, E. (1933). *The human problems of an industrial civilization*. New York: Macmillan.
- McDonough, J., & McDonough, S. (1997). *Research methods for English language instructors*. London: Arnold.
- McGroarty, M. E., & Zhu, W. (1997). Triangulation in classroom research: A study of peer revision. *Language Learning*, 47(1), 1–43.
- Mellow, J. D., Reeder, K., & Forster, F. (1996). Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition*, 18, 325–350.
- Melzi, G., & King, K. (2003). Spanish diminutives in mother–child conversations. *Journal of Child Language*, 30, 281–304.
- Mitchell, R., Parkinson, B., & Johnstone, R. (1981). *The foreign language classroom: An observational study*. Stirling Educational Monographs, No. 9. Stirling, Scotland: University of Stirling, Department of Education.
- Montrul, S. (2001). Agentive verbs of manner of motion in Spanish and English as second languages. *Studies in Second Language Acquisition*, 23, 171–206.
- Morrison, L. (1996). Talking about words: A study of French as a second language learners' lexical inferencing procedures. *The Canadian Modern Language Review*, 53, 41–75.
- Moskowitz, G. (1967). The FLint system: An observational tool for the foreign language class. In A. Simon & E. G. Boyer (Eds.), *Mirrors for behavior: An anthology of classroom observation instruments* (pp. 1–15). Philadelphia: Center for the Study of Teaching, Temple University.
- Moskowitz, G. (1970). *The foreign language instructor interacts*. Minneapolis: Association for Productive Teaching.
- Muchisky, D. (1983). Relationships between speech and reading among second language learners. *Language Learning*, 33, 77–102.
- Murphy, J. M. (1992). An etiquette for the nonsupervisory observation of L2 classrooms. *Foreign Language Annals*, 25(3), 215–225.
- Murphy, V. (1997). The effect of modality on a grammaticality judgment task. *Second Language Research*, 13, 34–65.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Retrieved February 12, 2004, from <http://ohsr.od.nih.gov/mpa/belmont.php3>

- National Institute for Health. (n.d.). *Human subject protections education for research teams*. Retrieved April 15, 2002, from <http://cme.nci.nih.gov/>
- Neufeld, A., Harrison, M., Hughes, K., Spitzer, D., & Stewart, M. (2001). Participation of immigrant women family caregivers in qualitative research. *Western Journal of Nursing Research*, 23, 575–591.
- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417–528.
- Norris, J. M., & Ortega, L. (2003). Defining and measuring L2 acquisition. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 717–761). New York: Blackwell.
- Nunan, D. (1989). *Understanding language classrooms: A guide for instructor initiated action*. New York: Prentice-Hall.
- Nunan, D. (1993). Action research in language education. In J. Edge & K. Richards (Eds.), *Instructors develop instructors research: Papers on classroom research and instructor development* (pp. 39–50). Oxford, UK: Heinemann.
- Nuremberg Code. (1949). From *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No. 10* (vol. 2, pp. 181–182). Washington, DC: U.S. Government Printing Office. Retrieved April 15, 2002, from <http://ohsr.od.nih.gov/nuremberg.php3>
- Office for Human Research Protections (OHRP) of the Department of Health and Human Services (DHHS). (n.d.). *IRB guidebook*. Retrieved April 15, 2002, from http://ohrp.osophs.dhhs.gov/irb/irb_guidebook.htm
- Office for Human Research Protections (OHRP) of the Department of Health and Human Services (DHHS). (n.d.). *Training modules for assurances: Investigator responsibilities and informed consent*. Retrieved April 15, 2002, from http://137.187.206.145/cbtng_ohrp/cbts/assurance/module2qset1_1.asp
- Office of Human Subjects Research (OHSR) of the National Institutes of Health. (2000, August). *Information sheet #5: Guidelines for writing research protocols*. Retrieved April 15, 2002, from http://ohsr.od.nih.gov/info/einfo_5.php3
- Ohta, A. S. (1995). Applying sociocultural theory to an analysis of learner discourse: Collaborative interaction in the zone of proximal development. *Issues in Applied Linguistics*, 6(2), 93–121.
- Oliver, R. (1998). Negotiation of meaning in child interactions. *The Modern Language Journal*, 82, 372–386.
- Oliver, R. (2000). Age differences in negotiation and feedback in classroom and pairwork. *Language Learning*, 50(1), 119–151.
- Oliver, R. (2002). The patterns of negotiation for meaning in child interactions. *The Modern Language Journal*, 86, 97–111.
- Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 109–148.
- Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139–162). New York: Russell Sage Foundation.

- Pavlenko, A., & Lantolf, J. P. (2000). Second language learning as participation and the (re)construction of selves. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 155–178). New York: Oxford University Press.
- Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Philips, S. U. (1972). Participant structures and communicative competence: Warm Indian Springs children in community and classroom. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. 370–394). New York: Teachers College Press.
- Philips, S. (1983). *The invisible culture: Communication in classroom and community on the Warm Springs Indian Reservation*. New York: Longman.
- Philp, J. (2003). Constraints on “noticing the gap”: Nonnative speakers’ noticing of recasts in NS–NNS interaction. *Studies in Second Language Acquisition*, 25, 99–126.
- Pica, T. (1984). The selective impact of classroom instruction on second language acquisition. *Applied Linguistics*, 6, 214–222.
- Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, 44, 493–527.
- Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. In G. Crookes & S. M. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 9–34). Clevedon, UK: Multilingual Matters.
- Pienemann, M., & Johnston, M., (1986). An acquisition-based procedure for second-language assessment. *Australian Review of Applied Linguistics*, 9, 92–122.
- Pintzuk, S. (1988). *VARBRUL programs* [Computer program]. Philadelphia: University of Pennsylvania, Department of Linguistics.
- Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47(1), 101–143.
- Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition*, 19, 499–508.
- Porte, G. K. (2002). *Appraising research in second language learning: A practical approach to critical analysis of quantitative research*. Amsterdam: John Benjamins.
- Portney, L. G., & Watkins, M. P. (1993). *Foundations of clinical research: Applications to practice*. Norwalk, CT: Appleton and Lange.
- Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing*, 10, 277–303.
- Ramanathan, V., & Atkinson, D. (1999). Ethnographic approaches and methods in L2 writing research: A critical guide and review. *Applied Linguistics*, 20(1), 44–70.
- Rand, D., & Sankoff, D. (1990). *GoldVarb: A variable rule application for the Macintosh (version 2.0)* [Computer program]. Montreal: Centre de Recherches Mathématiques, Université de Montréal.

- Reichardt, C. S., & Cook, T. D. (1979). Beyond qualitative versus quantitative methods. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 7–32). Beverly Hills, CA: Sage.
- Reynolds, D. (2001). Language in the balance: Lexical repetition as a function of topic, cultural background, and writing development. *Language Learning*, 51, 437–476.
- Roberts, M. (1995). Awareness and the efficacy of error correction. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Tech. Rep. No. 3, pp. 163–182). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.
- Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced and instructed conditions. *Studies in Second Language Acquisition*, 19(2), 223–247.
- Rodríguez, M., & Abreu, O. (2003). The stability of general foreign language classroom anxiety across English and French. *The Modern Language Journal*, 87, 365–374.
- Rounds, P. (1996). The classroom-based researcher as fieldworker: Strangers in a strange land. In J. Schachter & S. Gass (Eds.), *Second language classroom research: Issues and opportunities* (pp. 45–59). Mahwah, NJ: Lawrence Erlbaum Associates.
- Ruiyang, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, 22, 365–385.
- Rust, R. T., & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research*, 31, 1–14.
- Sachs, R., & Polio, C. (2004). *Reformulation as an alternative to error correction in second language writing*. Unpublished manuscript, Michigan State University, East Lansing.
- Samuda, V. (2001). Guiding relationships between form and meaning during task performance: The role of the instructor. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 119–140). London: Longman.
- Sasaki, Y. (1997). Material and presentation condition effects on sentence interpretation task performance: Methodological examinations of the competition experiment. *Second Language Research*, 13, 34–65.
- Schachter, J., & Gass, S. (1996). *Second language classroom research: Issues and opportunities*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26, 99–128.
- Schenkein, J. (1978). *Studies in the organization of conversational interaction*. New York: Academic Press.
- Schmidt, R. (1983). Interaction, acculturation and the acquisition of communicative competence. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and second language acquisition* (pp. 137–174). Rowley, MA: Newbury House.
- Schmidt, R., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn* (pp. 237–326). Rowley, MA: Newbury House.

- Schumann, F., & Schumann, J. (1977). Diary of a language learner: An introspective study of second language learning. In H. D. Brown, C. Yorio, & R. Crymes (Eds.), *Teaching and learning English as a second language: Trends in research and practice* (pp. 241–249). Washington, DC: TESOL.
- Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
- Scollon, R. (2001). Action and text: Toward an integrated understanding of the place of text in social (inter)action. In R. Wodak & M. Meyer (Eds.), *Methods in critical discourse analysis* (pp. 139–183). London: Sage.
- Shavelson, R. (1988). *Statistical reasoning for the behavioral sciences*. Boston: Allyn and Bacon.
- Silver, R. E. (1999). Input, output, and negotiation: Conditions for second language development. B. Swierzbinska, F. Morris, M. E. Anderson, C. A. Klee, & E. Tarone (Eds.), *Social and cognitive factors in second language acquisition: Selected proceedings of the 1999 Second Language Research Forum* (pp. 345–371). Somerville, MA: Cascadia.
- Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. London: Oxford University Press.
- Skehan, P. (1998). *A cognitive approach to language learning*. New York: Oxford University Press.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49, 93–120.
- Spada, N. (1994). Classroom interaction analysis. *TESOL Quarterly*, 28(4), 685–688.
- Spada, N. (1997). Form-focused instruction and second language acquisition: A review of classroom and laboratory research [State of the Art Article]. *Language Teaching*, 30(2), 73–87.
- Spada, N., & Fröhlich, M. (1995). *The Communicative Orientation of Language Teaching (COLT) observation scheme: Coding conventions and applications*. Sydney, Australia: Macquarie University, National Centre for English Language Teaching and Research.
- Spada, N., & Lightbown, P. M. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition*, 15(2), 205–224.
- Staphorsius, G., & Verhelst, N. D. (1997). Indexering van de leesteknik [The development of a domain-referenced index of the decoding load of texts]. *Pedagogische Studiën*, 74, 154–164.
- Storch, N., & Tapper, J. (1996). Patterns of NNS student annotations in identifying areas of concern in their writing. *System*, 24(3), 323–336.
- Sullivan, N., & Pratt, E. (1996). A comparative study of two ESL writing environments: A computer-assisted classroom and a traditional oral classroom. *System*, 29(4), 491–501.
- Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, 82, 320–337.

- Swain, M., & Lapkin, S. (2001). Focus on form through collaborative dialogue: Exploring task effects. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning and testing* (pp.99–118). Harlow, UK: Longman.
- Tetnowski, J., & Damico, J. (2001). A demonstration of the advantages of qualitative methodologies in stuttering research. *Journal of Fluency Disorders*, 26, 17–42.
- Thompson, R., & Jackson, S. (1998). Ethical dimensions of child memory research. *Applied Cognitive Psychology*, 12, 218–224.
- Thorndike, E., & Lorge, I. (1944). *The teachers word book of 30,000 words*. New York: Columbia University.
- Toth, P. D. (2000). The interaction of instruction and learner-internal factors in the acquisition of L2 morphosyntax. *Studies in Second Language Acquisition*, 22(2), 169–208.
- Tse, L. (1996). If you lead horses to water they will drink: Introducing second language adults to books in English. *California Reader* 29(2), 14–17.
- Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly*, 26, 713–730.
- Ullman, R., & Geva, E. (1983). *Classroom observation in the L2 setting: A dimension of program evaluation*. Toronto: Modern Language Centre, Institute for Studies in Education.
- Ullmann, R., & Geva, E. (1985). Expanding our evaluation perspective: What can classroom observation tell us about Core French programs? *The Canadian Modern Language Review*, 42(2), 307–323.
- University of Pennsylvania. (n.d.). *Institutional review boards standard operating policies: Section 500 SC reviews requiring special considerations* (pp. 3–4). Retrieved January 24, 2004, from <http://www.upenn.edu/regulatoryaffairs/human/SOP/SC-501.pdf>
- Valdman, A. (1993). Replication study. *Studies in Second Language Acquisition*, 15, 505.
- Van Daalen-Kaptein, M., Elshout-Mohr, M., & de Glopper, K. (2001). The acquisition of word meanings as a cognitive learning process. *Journal of Verbal Learning and Verbal Behavior*, 20, 386–399.
- van der Veer, R., van Ijzendoorn, M., & Valsiner, J. (1994). General introduction. In R. van der Veer, M. van Ijzendoorn, & J. Valsiner (Eds.), *Reconstructing the mind: Replicability in research on human development* (pp. 1–10). Norwood, NJ: Ablex.
- VanPatten, B., & Cadierno, T. (1993). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, 77(1), 45–56.
- van Someren, M., Barnard, Y., & Sandberg, J. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.
- Verburg, M., & Huijgen, M. (1994). *Van dale Juniorwoordenboek Nederlands* [Van Dale junior dictionary of Dutch]. Utrecht: Van Dale Lexicografie.
- Wallace, M. J. (1998). *Action research for language instructors*. Cambridge, UK: Cambridge University Press.
- Watson-Gegeo, K. (1988). Ethnography in ESL: Defining the essentials. *TESOL Quarterly*, 22, 575–592.

- Watson-Gegeo, K. (1997). Classroom ethnography. In N. H. Hornberger & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 8. Research methods in language and education* (pp. 135–144). Dordrecht, The Netherlands: Kluwer.
- White, L. (1985). The “Pro-drop” parameter in adult second language acquisition. *Language Learning*, 35, 47–62.
- White, L., Spada, N., Lightbown, P. M., & Ranta, L. (1991). Input enhancement and L2 question formation. *Applied Linguistics*, 12(4), 416–432.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Willett, J. (1995). Becoming first graders in an L2: An ethnographic study of L2 socialization. *TESOL Quarterly*, 29, 473–503.
- Williams, J. (1999). Learner-generated attention to form. *Language Learning*, 49, 583–625.
- Williams, J., & Evans, J. (1998). What kind of focus and on which forms? In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 139–155). Cambridge, UK: Cambridge University Press.
- Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies*. Cambridge, UK: Cambridge University Press.
- World Medical Association (2000). *Declaration of Helsinki. Ethical principles for medical research involving human subjects*. Retrieved April 15, 2002, from <http://www.wma.net/e/policy/b3.htm>
- Wray, A. (2001). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.
- Young, R. (1991). *Variation in interlanguage morphology*. New York: Peter Lang.
- Young, R., & Bayley, R. (1996). VARBRUL analysis for second language research. In R. Bayley & D. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 253–306). Amsterdam: John Benjamins.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24, 1–27.
- Zsiga, L. (2003). Articulatory timing in a second language. *Studies in Second Language Acquisition*, 25(3), 399–432.

This page intentionally left blank

Author Index

A

- Abreu, O., 260, 261, 382
Adair, G., 187, 370
Adams, R., 238, 240, 370
Akiyama, Y., 59n, 370
Allen, J. P. B., 190, 191, 370
Allinson, D., 304, 382
Allwright, R., 190, 201, 217, 370
Alvarez-Torres, M., 8, 52, 57, 89, 142, 268, 375
The American Heritage Dictionary of the English Language, Fourth Edition, 1, 370
American Psychological Association, 6n, 9n, 21, 124, 126, 283, 319, 370
Atkinson, D., 168, 368

B

- Bailey, K. M., 178, 190, 203, 204, 205, 217, 370, 371
Barcroft, J., 265, 266, 371
Bard, E., 56, 57, 371
Bardovi-Harlig, K., 86, 92, 232, 371
Basturkmen, H., 214, 374
Bates, E., 61, 371
Baumrind, 31, 371
Bayley, R., 292, 385
Beck, M. L., 61, 371
Beebe, L., 89, 371
Berg, E. C., 299, 301, 302, 371
Bergthold, B., 204, 371
Bernard, H. R., 230, 371

- Birdsong, D., 53, 371
Bley-Vroman, R., 52, 53, 54, 61, 64, 372
Blok, H., 10, 11, 12, 374
Braunstein, B., 204, 371
Breen, M. P., 94, 95, 372
British Association for Applied Linguistics, 25n, 372
Brown, J. A. C., 187, 372
Brown, J. D., 92, 164, 167, 372
Brown, R., 232, 372
Butler, C., 255, 372

C

- Cadierno, T., 213, 384
Carmines, E., 244, 372
Carpenter, P. A., 63, 377
Carson, J., 168, 372
Chalmers, D., 25n, 372
Chan, C. Y., 51, 54, 376
Chaudron, C., 164, 166, 191, 216, 244, 372
Chomsky, N., 48, 372
Clément, R., 145, 373
Cohen, A., 77, 373
Cohen, J., 243, 373
Cohen, J. D., 63n, 373
Cooil, B., 244, 382
Cook, T. D., 2, 382
Coulthard, R. M., 190, 383
Crookes, G., 216, 244, 370, 373

D

- Damico, J., 163, 384

Davis, K. A., 180, 373
 de Gopper, K., 10, 12, 374, 384
 de Graaf, R., 288, 373
 de Groot, A., 258, 378
 De Guerrero, M., 223, 373
 Denzin, N., 241, 373
 Department of Health and Human Services
 (DHHS), 34, 373
 Derwing, T., 134, 373
 Dörnyei, Z., 92, 94, 145, 371, 373
 Duff, P., 28, 168, 373
 Duffield, N., 61, 258, 373, 374

E

Early, M., 28, 373
 Egi, T., 85, 374
 Ellis, N., 283, 374
 Ellis, R., 12, 74, 88n, 171, 198, 212, 214, 240,
 299, 300, 374, 385
 Elshout-Mohr, M., 11, 384
 Endnote (Version 7.0), 316, 374
 Eubank, L., 61, 374
 Evans, J., 214, 385

F

Falodun, J., 65, 71n, 381
 Fanselow, J. F., 190, 374
 Felix, S., 53, 64, 372
 Fernández-García, M., 57, 58, 89, 142, 268,
 375
 Fiksdal, S., 86, 374
 Flatt, M., 63n, 373
 Fleischman, N. J., 204, 371
 Fletcher, P., 251, 256, 385
 Forster, F., 153, 154, 155, 188, 379
 Foster, P., 89, 161, 220, 372, 383
 Fotos, S., 74, 374
 Fraenkel, J., 124, 180, 374
 Francis, W., 116, 374
 Fröhlich, M., 190, 191, 192, 193, 196, 370, 383
 Frota, S. N., 177, 382
 Fujii, A., 202, 203, 378
 Fukkink, R., 10, 11, 12, 374

G

Gaies, S. J., 232, 374

Gass, S., 3, 8, 9, 21, 22, 41, 43, 44, 44n, 49, 50,
 52, 57, 58, 60, 66, 77, 78, 85, 89, 91,
 114, 140, 141, 142, 151, 157, 203,
 206, 224, 268, 307, 375, 378, 381,
 382
 Geertz, C., 3, 375
 Georgetown University, 326, 375
 Geva, E., 190, 191, 193-195, 384
 Glew, M., 58, 59, 375
 Goldschmidt, M., 92, 375
 Goldstein, T., 168, 375
 Granger, S., 97, 375
 Grotjahn, R., 2, 4, 376
 Gwet, K., 244, 376

H

Hall, A., 174, 376
 Harrington, M., 62, 64, 374, 377
 Harrison, M., 26, 380
 Hartford, B. S., 86, 371
 Harvard University, 26, 376
 Hawkins, R., 51, 54, 376
 He, X., 299, 300, 374
 Heath, A. W., 305, 376
 Heath, S. B., 168, 376
 Helms-Park, R., 18, 376
 Hendrick, C., 22, 376
 Hird, B., 94, 95, 372
 Holbrook, M. P., 204, 371
 Houck, N., 91, 224, 375
 Hughes, A., 251, 256, 385
 Hughes, K., 28, 380
 Hulstijn, J. H., 186, 376
 Hunt, K. W., 231, 376

I

Inagaki, S., 54, 376
 Ionin, T., 51, 53, 376
 Ioup, G., 53, 54, 64, 372

J

Jackson, S., 32, 34, 384
 Jiang, N., 15, 376
 Jick, T., 183, 376
 Johnson, D. M., 167, 168, 172, 181, 216, 219,
 377
 Johnson, J., 51, 377

Johnston, M., 234, 235, 236, 381
 Johnstone, R., 190, 379
 Joo, H-R., 52, 372
 Juffs, A., 64, 377
 Just, M. A., 63, 377

K

Kanagy, R., 65, 71n, 381
 Kasper, G., 86, 377
 Kennedy, E., 153, 377
 King, K. A., 168, 299, 303, 377, 379
 Kucera, H., 116, 374

L

Labov, W., 176, 377
 Lakshmanan, U., 20, 21, 59n, 377
 Lantolf, J. P., 168, 381
 Lapkin, S., 74, 129, 182, 379, 384
 Lemelin, S., 141, 375
 Leow, R., 11, 12, 299, 301, 302, 303, 377
 Lieven, E. V. M., 303, 377
 Lightbown, P. M., 192, 214, 234, 235, 378,
 383, 385
 Lincoln, Y., 241, 373
 Loewen, S., 214, 372
 Long, M., 3, 244, 372, 378
 Lorge, I., 116, 384
 Lotto, L., 258, 376
 Lynch, B. K., 191, 378
 Lyster, R., 193, 214, 237, 239, 378

M

Macaro, E., 75, 76, 84, 378
 Macias, J., 168, 378
 Mackey, A., 41, 44, 57, 58, 77, 78, 85, 89, 142,
 157, 198, 202, 203, 234, 236, 243,
 268, 307, 375, 378
 MacWhinney, B., 61, 63n, 97, 233, 369, 371,
 379
 Markee, N., 222, 233, 379
 Mason, J., 162, 175, 379
 Masterson, D., 61, 372
 Mayo, E., 187, 379
 McDonough, J., 191, 217, 379
 McDonough, K., 78, 202, 203, 378
 McDonough, S., 191, 217, 378
 McGroarty, M. E., 181, 379

Mellow, J. D., 153, 154, 155, 188, 379
 Melzi, G., 299, 303, 379
 Milton, M., 94, 95, 372
 Mitchell, R., 190, 379
 Montrul, S., 52, 54, 379
 Morgan-Short, K., 11, 12, 377
 Morrison, L., 82, 379
 Moskowitz, G., 190, 379
 Muchisky, D., 18, 379
 Munro, M., 134, 373
 Murphy, J. M., 189, 379
 Murphy, V., 51, 379

N

National Commission for the Protection of
 Human Subjects of Biomedical
 and Behavioral Research, 26, 27,
 31, 39, 379
 National Institute for Health, 26, 380
 Nelson, G. L., 168, 372
 Neufeld, A., 28, 380
 Newport, E., 51, 377
 Norris, J. M., 215, 242, 283, 380
 Nunan, D., 190, 191, 192, 216, 219, 380
 Nuremberg Code, 38, 380

O

Office for Human Research Protections
 (OHRP) of the Department of
 Health and Human Services
 (DHHS), 26, 37, 380
 Office of Human Subjects Research
 (OHSR) of the National Insti-
 tutes of Health, 26, 380
 Ohta, A. S., 172, 380
 Oliver, R., 94, 95, 198, 236, 243, 247, 370, 376,
 380
 Ortega, L., 215, 242, 265, 283, 380
 Orwin, R. G., 221, 380

P

Parkinson, B., 190, 379
 Pavlenko, A., 168, 381
 Pedhazur, E., 244, 381
 Philips, S. U., 168, 381
 Philp, J., 3, 13, 14, 85, 234, 236, 245, 299, 300,
 303, 378, 381

Pica, T., 3, 65, 71n, 233, 379
 Pienemann, M., 234, 235, 236, 381
 Pintzuk, S., 291, 381
 Polio, C., 21, 22, 232, 238, 240, 381
 Porte, G. K., 21, 381
 Portney, L. G., 244, 381
 Pratt, E., 164, 383
 Prévost, P., 61, 373
 Provost, J., 63n, 373

Q

Qi, D. S., 182, 381

R

Ramanathan, V., 167, 168, 170, 370, 381
 Rand, D., 291, 381
 Ranta, L., 193, 214, 237, 239, 378, 385
 Reeder, K., 153, 154, 155, 188, 379
 Reichardt, C. S., 2, 382
 Reynolds, D., 17, 382
 Roberts, M., 203, 382
 Robertson, D., 56, 57, 371
 Robinson, P., 261, 262, 382
 Rodríguez, M., 260, 261, 382
 Rose, K., 86, 377
 Ross-Feldman, L., 157, 375
 Rounds, P., 30, 382
 Ruiyang, Y., 304, 382
 Rust, R. T., 244, 382

S

Sachs, R., 238, 240, 382
 Samuda, V., 198, 382
 Sankoff, D., 291, 381
 Sasaki, Y., 62, 382
 Schachter, J., 206, 382
 Schegloff, E. A., 234, 382
 Schenkein, J., 342-344, 382
 Schmidt, R., 171, 177, 382
 Schumann, F., 177, 383
 Schumann, J., 177, 383
 Schütze, C., 50, 383
 Scollon, R., 164, 383
 Selinker, L., 44n, 49, 375
 Shavelson, R., 268, 383
 Silver, R. E., 234, 383
 Sinclair, J. M., 190, 383

Skehan, P., 89, 161, 240, 383
 Sorace, A., 56, 57, 371
 Spada, N., 190, 192, 193, 196, 200, 213, 214,
 234, 235, 370, 378, 383, 385
 Spitzer, D., 28, 380
 Staphorsius, G., 10, 383
 Stewart, M., 28, 380
 Storch, N., 251, 252, 383
 Sullivan, N., 164, 383
 Svetics, I., 141, 375
 Swain, M., 74, 129, 383, 384

T

Takahashi, T., 89, 371
 Tapper, J., 251, 252, 383
 Task Force on Statistical Inference 283, 385
 Tatsumi, T., 202, 203, 378
 Teranishi, K., 59n, 377
 Tetnowski, J., 163, 384
 Thompson, R., 32, 34, 384
 Thorndike, E., 116, 384
 Thwaite, A., 94, 95, 372
 Toth, P. D., 255, 384
 Tse, L., 218, 384
 Tyler, A., 86, 384

U

Uliss-Weltz, R., 90, 371
 Ullman, R., 190, 384
 University of Pennsylvania, 34, 384

V

Valdman, A., 21, 22, 384
 Valsiner, J., 22, 384
 Van Daalen-Kapteijns, M., 11, 384
 van der Veer, R., 22, 384
 van Ijzendoorn, M., 22, 384
 VanPatten, B., 213, 384
 van Someren, M., 79, 384
 Varonis, E., 3, 66, 375
 Verburg, M., 10, 384
 Verhelst, N. D., 10, 383
 Villamil, O., 223, 373

W

Wallace, M. J., 216, 384

Wallen, N., 124, 180, 374
Watkins, M. P., 244, 381
Watson-Gegeo, K., 168, 179, 242, 384, 385
Wexler, K., 51, 53, 376
White, L., 61, 192, 258, 373, 374, 385
Wiebe, G., 134, 373
Wilkinson, L., 283, 385
Willett, J., 3, 298, 302, 303, 385
Williams, J., 7, 8, 198, 214, 299, 300, 301, 385
Woods, A., 251, 256, 385
Wooley, J. D., 63, 377
World Medical Association, 26, 39, 385
Wray, A., 172, 173, 385

Y

Young, R., 291, 292, 385
Yuan, F., 12, 240, 372, 385

Z

Zambo, L. J., 204, 371
Zeller, R., 244, 373
Zhu, W., 181, 379
Zsiga, L., 256, 385

This page intentionally left blank

Subject Index

NOTE: Locators in italics refer to the Glossary.

A

Abstract representations, 48–49
Abstracts, research report, 5, 7, 310, 318, 320, 350
Academic English word frequency index, 116
Academic Word List, 116
Acceptability judgments, 45–46, 48–55, 151, 350
Accuracy, operationalization of, 240
Action research, 212, 216–219, 350
Advising sessions, 86
Age of participants, 106, 126, 229
Alpha level, 267–268, 270
Analysis, research, *see* Data analysis
Analysis of covariance, *see* ANCOVA
Analysis of variance, *see* ANOVA
ANCOVA (analysis of covariance), 276, 350
Anonymity, 28, 29, 75, 123, 124, 126
 in reports, 311, 312
 see also Confidentiality
ANOVA (analysis of variance), 265, 282, 274–275, 280, 291, 348, 350
 formula, 347
 repeated measures, 266, 277, 280
 two-way, 275–276
Anxiety, 260–261
Appropriateness of task, 139–140
Artificiality of experimental data, 91, 96
Association, strength of, 137, 282, 366
Associational research, 137–138, 145–146, 350
 see also Correlational research

Attention, 75, 81, 110, 112–113, 141, 213
Attitudes, participants', 107, 114, 118, 202
Average, arithmetical, *see* Mean

B

Bell curve (normal distribution), 261–263, 351, 361
Belmont Report, 26, 27, 39–40
Beneficence, 40
Between-groups studies, 146–148
Biodata, 124–127, 128, 351
Boredom, participant, 114, 118
Brown Corpus, 116
Budget, justification in grant proposals, 319

C

CALL (computer-assisted language learning), 246
Case studies, 171–173, 305–306, 350
Categorical data, 278
Central tendency, measures of, 228, 254–258, 307, 360
 see also Mean; Median; Mode
Charts, 13, 14, 291
CHAT coding system, 97, 233–234
Chi square, 278–279, 291, 348, 351
 distribution of, 281–282
CHILDES corpus, 97, 233–234, 351
Children
 child-directed speech, 102
 effectiveness of instruction, 214–215
 formulaic sequences and SLA, 172–173
 informed consent, 32–34

- maturation, 115
 requests, ESL, 171–172
 Universal Grammar, 48
 verb inflection and pronoun use, 20–21
- Circumlocution, 76
- Citation indexes, 18
- Class, social, 168–169
- Classroom, research, 185–220, 349
 - action research, 212, 216–219, 349
 - analysis of data, 215
 - contexts, 185–186, 211, 215
 - control groups, 214–215
 - data segmentation, 210–211
 - debriefing, 209–210
 - disruption, 187, 188–189, 210
 - ethnographies, 212
 - intervening variables, 215
 - introspective methods, 201–205
 - see also* Diaries/journals; Stimulated recall; Uptake sheets
 - logistics, 206–209
 - meta-analyses, 215
 - multiple methods, 186, 206, 219–220, 306
 - practical considerations, 205–212
 - purposes and types of research, 212–219
 - research reports, 190, 306
 - transcription conventions, 342–344, 345–346
 - see also* Instruction, effect of; Observations, classroom; Teachers; *and under* Coding; Confidentiality; Consent; Qualitative research; Recordings; Triangulation
- Clinical science research, 38–39, 244–245
- Clubs, language exchange, 176
- CMC (computer-mediated communication), 75, 352
- Coding, 221–249, 350
 - classroom data, 190–201, 210–211, 215, 237–238, 239
 - common systems and categories, 230, 231–234
 - see also* CHAT; SOC; T-units
 - comparability, 230, 242
 - custom-made systems, 201, 234–241
 - generalizability, 234
 - highest level of measurement, 230
 - interval data, 229
 - mechanics, 246–248
 - nominal data, 226–227
 - open, 241
 - oral data, 222–225
 - ordinal data, 227–228
 - pilot studies, 248
 - preparation of data, 221–225
 - proportion of data to code, 247
 - qualitative data, 233–234, 241–242
 - rater reliability, 242–246, 313, 349, 358
 - replication studies, 230
 - research reports on, 245–246, 313
 - segmentation of data, 210–211, 247, 354
 - systems, 230–242
 - timing of coding decisions, 248
- Cognitive processes, 75–85
- Cohen's d , 283, 349
- Cohen's κ , 243–244, 245
- Collaborative research, *see* Action research and Qualitative Research
- Collection, *see* Data collection
- COLT (Communicative Orientation of Language Teaching) scheme, 191–192, 193, 196–197, 200
- Commentary, learner, 75, 76, 77
- Communicative Orientation of Language Teaching scheme, *see* COLT
- Comparison group design, 146–147, 157, 351
- Compensation of participants, 34, 211, 318
- Competence, linguistic and pragmatic, 171
- Competition Model, 61–62
- Complexity, operationalization of, 240
- Computer technology
 - computer-assisted language learning, 246
 - computer-mediated communication, 75, 351
 - computer-mediated research, 63–64, 75, 76, 109, 164, 246
 - see also* Software
- Confidentiality, 27, 28–29, 37
 - in classroom research, 210
 - in debriefing, 318
 - in research reports, 311, 312
 - see also* Anonymity
- Confirmability, 180, 352
- Consciousness-raising tasks, 74, 352
- Consensus tasks, 72–74, 352
- Consent, informed, 26–36, 358
 - child participants, 32–34
 - classroom research, 188–189, 207, 208, 209

- codes of research ethics, 39, 40
 comprehension as condition, 31–34
 local norms, 35
 forms, 31–32, 33, 34, 35–36, 311,
 322–325
 immigrant and refugee populations,
 28, 29
 and recordings, 207
 suppliancy of sufficient information,
 27–31
 trust, problems of, 28, 29, 35
 voluntary participation, 34–35
 withholding of information, 30–31,
 176
see also Institutional review
- Consistency, internal, 130
- Consonants, word-final, 102, 103–104, 108
- Constraints, practical, 308
- Construct validity, 107–108, 352
- Consultants, acknowledgment of, 314,
 317–318
- Contact information, researcher's, 27, 28
- Contexts
 - classroom research, 185–186, 211, 215
 - limitations, 303
 - qualitative research, 172, 180
- Contingency plans, *see also* Protocols, 311
- Control groups, 115, 214–215, 219
 - control group design, 148, 157, 352
- Control variables, 104–105, 352
- Convenience sample, 122, 128, 352
- Corpora, 29, 97–98, 352
- Correlation coefficient, 129n, 352
- Correlational data analysis, 284–291
 - see also* Factor analysis; Kendall Tau
 test; Pearson product-moment correla-
 tion; Regression; Spearman Rho calcula-
 tion
- Correlational research, 124, 137, 145–146,
 352
 - see also* Associational research
- Cost/benefit analyses, 138
- Counterbalancing, 143–145, 352
- Covariance, 353
 - analysis of (ANCOVA), 276, 350
- Credibility, 179–180, 353
- Criterion-related validity, 108, 353
- Critical discourse analysis, 163–164
- Critical research, 163–164
- Critical value, 270, 280–281, 353
- Cronbach's α , 130, 353
- Crossword puzzles, 81–82
- Cues and cue strengths in Competition
 Model, 61–62
- Cyclical processes, 163
 - data analysis, 178–179, 182, 353
- D**
- Data, 353
 - Naturalistic, 86, 361
 - quantitative/qualitative distinctions,
 165
- Data analysis, 250–296
 - classroom research, 215
 - credibility, transferability,
 confirmability, and dependability,
 179–181
 - cyclical, 178–179, 182, 353
 - diaries/journals, 178
 - grounded theory, 179, 357
 - inductive, 179, 357
 - interpretative vs. statistical, 2, 4
 - meta-analyses, 283–284
 - qualitative research, 2, 4, 178–182
 - quantitative research, 2, 4, 250–296
 - research reports, 6, 12–13, 306,
 312–314
 - triangulation, 181–182
 - see also* Correlational data analysis; Sta-
 tistics
- Data collection, 43–99, 354
 - collector characteristics and validity,
 115, 118
 - computer-mediated, 63–64, 75, 76,
 109, 164, 246
 - experimental/quasi-experimental vs.
 nonexperimental, 2–3, 4
 - feasibility assessment, 19, 43–44
 - formal models of language, 48–61
 - interaction-based research, 46–47,
 65–75
 - introspective measures, 75–76, 77–85
 - measures, *see* Acceptability judgments;
 Case studies; Consciousness-raising
 tasks; Consensus tasks; Diaries/jour-
 nals; Discourse completion tests;
 Ethnographies; Imitation, elicited; Im-
 mediate recall; Interpretation tasks; In-
 terviews; Introspective measures;
 Jigsaw tasks; Magnitude estimation;
 Moving window; Narratives, elicited;
 Naturalistic settings; Observations;
 Online tasks; Picture description tasks;
 Questionnaires; Reaction time; Role

- plays; Sentence interpretation; Sentence matching; Spot the difference; Stimulated recall; Think-alouds; Truth-value judgments; Video playback for interpretation
 physical environment for, 115, 118
 pilot testing, 43–44
 processing research, 61–64
 quantitative/qualitative distinctions, 165–166
 research reports on, 306, 312
 significance of measures, 44–48
 sociolinguistic/pragmatics-based research, 47–48, 85–92
 strategy-based research, 75–85
 and theoretical framework, 44–48
see also Corpora; Ethical issues; Institutional review; Participants; *and under*
 Qualitative research
- Data elicitation, *see* Data collection
- Data sampling/segmentation, 210–211, 247, 354
- Data storage logistics, 311
- Databases
 online search, 310
see also Corpora
- Dative alternation, 74
- DCT, *see* Discourse completion tests
- Debriefing, 30, 189–190, 209–210, 318, 354
- Deduction, reasoning by, 76
- Department of Health and Social Services, U.S., 26, 37
- Dependability, 180–181, 354
- Description, rich or thick, 162, 180, 306–307, 367
- Descriptive research, 124, 167
- Design of qualitative studies, 2, 167–182
 data collection, 165–166, 167–178
see also Case studies; Diaries/journals; Ethnographies; Interviews; Observation
 hypotheses, 164, 178–179, 182
 multiple methods, 170
 and quantitative research, 2–5
 combination, 2–3, 4, 164, 182, 307
 differences, 2, 5, 16, 93, 96, 165–166
 questionnaires, 93, 96
 researcher's participation, 170, 176
 sociocultural context as focus, 168
 statistics with qualitative data, 307
see also under Coding; Contexts; Data analysis; Triangulation
- Design of quantitative studies, 137–161
 counterbalancing, 143–145, 353
 finalizing project, 158–159
 group assignment, 143–145, 157
 intact classes, 141–143, 150, 358
 measuring effect of treatment, 148–150, 158
 reports on, 311
 research materials, 138–141
 types of design, 145–158
 comparison group, 146–147, 157, 351
 control group, 148, 157, 352
 correlational, 124, 137, 145–146, 352
 experimental and quasi-experimental, 146–148, 157, 356, 363
 factorial, 151–152, 158, 356
 one-shot, 156–157, 158, 361
 posttest only, 149–150, 158, 363
 pretest/posttest, 148–149, 158, 363
 repeated measures, 150–151, 158, 364
 time-series, 152–155, 158, 188, 367
- Developmental stages, Piennemann-Johnston hierarchy, 234–236
- Deviation, standard, 259–261, 262, 307, 366
- DHHS (U.S. Department of Health and Social Services), 26, 37
- Diaries/journals, 75, 169, 176–178, 354
 in classroom research, 201, 203–205, 217–218
- Dictionary use, learners', 76
- Dictogloss tasks, 67, 74
- Difference
 spot the difference tasks, 67–71, 140–141
 standard error of (SED), 269–270, 347, 366
- Digital data, 206, 225
- Disclosure of goals, partial and full, 30–31, 50, 117, 119, 149, 176
- Discourse completion test (DCT), 47–48, 89–91, 93, 353
- Discourse markers, 76
- Dispersion, measures of, 258–261, 360
- Distance learning, 75
- Distractor questions and items, 50
- Distribution, 354
 normal, 261–263, 360
- DMDX software, 246
- Dropout rate, *see* Mortality, participant
- Duncan's multiple range test, 275, 355

- Dyads, 355
 Dynamic research design, 2, 169
- E**
- E-Prime* computer software, 63n
 Educational/home community comparisons, 168–169
 Effect
 main, 359
 size of, 282–283, 283–284, 348, 355
 Elicited imitation, 46, 55–56, 355
 Elicited narratives, 87–89, 355
 Emic perspectives, 163, 165, 168, 169, 170, 355
 Empirical research, 355
 Emotional variables, 177
 Endnote (ISI ResearchSoft) software, 316
 Equivalence of forms, and reliability, 130
 Eta², 282, 348, 355
 Ethical issues, 25–36
 and corpora, 29
 codes, 39–40
 and data gathering, 25–36
 guidelines, U.S., 25–36
 medical experimentation, 38–39
 participant observation, 176
 research reports on, 308, 311, 312
 see also Anonymity; Confidentiality; Consent, informed; Institutional review
 Ethnographies, 165–166, 167–171, 212, 355
 Etic perspectives, 163, 165, 356
 Etiquette, research, 314, 317–318
 in classroom settings, 187, 188–189, 190
 Experimental research, 2–3, 4, 137–138, 356
 design types, 146–148, 157, 356
 quasi-experimental, 146, 363
 sampling, 124, 146
 Extra-experimental factors, 114–15, 118
 Eye movements, 246
- F**
- Face validity, 107, 143, 356
 Factor analysis, 290–291, 356
 Factorial research design, 151–152, 158, 356
 Fatigue, participant, 50–51, 62, 114, 118
 Feasibility assessment, 18–19, 43–44
 Feedback
 classroom observations, 198, 199
 negative, 236–237
 post-research, to participants, 318
 reformulation vs. error correction, 182
 response generated by, 214
 type as research variable, 102, 103, 104
 variables affecting study, 102
 written, processing of, 182
 Filler or distractor questions, 50
 Film, in elicited narrative studies, 88–89
 Finalizing research, checklist for, 158–159
 First language background, 104–105, 109–110, 118, 126
 ambiguity of term, 126–127
 Fisher's exact test, 279, 356
 Fluency, operationalization of, 240
 Focus groups, 173, 356
 Form, focus on, 213–214
 Formal models of language, 48–61
 Forms
 biodata collection, 125–126
 informed consent, 31–32, 33, 34, 35–36, 322–325
 institutional review application sample, 326–341
 continuing review sample, 333–334
 exemption of review sample, 338–341
 expedited review sample, 335–337
 full/complete review sample, 326–332
 Formulæ
 linguistic, 171, 172–173
 statistical, 347–349
 Freedom, degrees of, 270, 280–281, 354
 Frequency
 data frequency distribution, 307
 polygons (line graphs), 251, 256
 statistical measures of, 251–254, 360
 word, indexes of, 116
 Friedman test, 280, 356
 Funding, 37, 309, 318, 319
 Future research topics, 15, 17–18, 299, 302–304, 314
- G**
- Gender
 gender-neutral writing in reports, 319
 of participants, 103–104, 104–105, 106, 120, 126
 Generalizability, 2, 355
 common coding systems and, 234

control variables and, 105
 laboratory research to other settings, 186
 qualitative research, 2, 163, 171, 172–173, 173–174, 178
 and replication, 22, 124
 research reports on, 302, 303
 sample size and, 123–124
 and validity, 119, 123–124
 Goals of study, avoiding giving away, 30, 50, 117, 119, 149
 Grammaticality judgment, *see also* Acceptability judgment, 49n
 Grant applications, 37, 309, 318, 319
 Graphic representation of data, 13, 14, 317, 318
 correlational, 285–287, 289
 descriptive, 251, 252–254, 255–256, 307
 SPSS statistical package, 291
 Grid-based schemes, 94, 95
 Grounded theory, 179, 357
 Group behavior, 167–171

H

Haberman's residuals, 279
 Halo effect, 114, 118, 174, 357
 Handedness, 9, 109
 Hard and soft data, 2
 Hawthorne effect, 114, 118, 155, 176, 187–188, 357
Helsinki, Declaration of, 26, 39
 Heritage speakers, 120, 169–170
 Holistic perspective, 163, 165, 168, 169, 171
 Home/educational community comparison, 168–169
 Human subjects committees, *see* Institutional review (Boards)
 Hypotheses, 1, 19–21, 100–101, 166, 357
 action research, 218
 directional, one-way, 101, 354, 361
 formation, 93, 100–101, 155, 178–179, 182
 non-directional, two-way, 101, 361, 368
 null, 100–101, 265–267
 one-tailed versus two-tailed, 270–271, 281
 in qualitative research, 164, 178–179, 182
 research reports and, 298, 309
 statistics and, 100–101, 281

I

Ideology, 163–164
 Imitation, elicited, 46, 55–56, 355
 Immediate recall, 85
 Immigrant populations, 28, 29, 86, 115
 Impartiality, 163–164
 Implications of results, 301–302, 304, 314
 Individual learner, focus on, 171, 172
 Inductive data analysis, 179, 357
 Inferential statistics, 269–280
 Informant checking, 165
 Information, and consent, 27–31, 358
 Incomplete disclosure, 30–31, 176
 Information-exchange tasks, 357
 Information-gap tasks, 358
 Informed consent committees, *see* Institutional review (Boards)
 Innateness, 48
 Input, 147, 213
 incomplete, impoverished, 48
 length of exposure to, 102, 103–104, 106, 107, 108
 modified, 66
 standardized, 66
 Institutional review, 36–41
 application forms, 326–341
 Boards (IRBs), 26–27, 29, 37–38, 40–41, 44, 358
 continuing applications, 38, 333–334
 exemptions, 37, 38, 338–341
 expedited applications, 37–38, 335–337
 full applications, 36–37, 326–332
 reasons for development, 38–39
 Instruction, effect of, 153–155, 212, 213–215, 355
 explicit and implicit, 110
 Instructions to test participants, 117–118, 119, 140
 Instructors, classroom, *see* Teachers
 Instrument reliability, 129–130, 219, 312, 358
 Instrumentation effects, 116–119
 Insufficient examples, 140–141
 Insufficient tokens, 139
 Intact class research design, 141–143, 150, 358
 Interaction, 3, 172, 303–304, 358
 Interaction-based research
 coding of data, 237–238, 239
 collection of data, 44, 46–47, 65–75
 Internet, 18, 29, 97

Interpretation tasks, 58–59
 Interpretative analysis, 2–3, 4
 Interval scales, 229, 271, 288, 358
 Interviews, 169, 173–175, 358
 open-ended, 170
 semi-structured, 173, 365
 standardized (structured), 173, 366
 unstructured, 173, 366
 Introspective data collection measures,
 75–85, 201–205, 359
 see also Diaries/journals; Recall;
 Think-alouds; Uptake sheets; Verbal
 reporting
 IRBs, *see* Institutional review (Boards)

J

Jigsaw tasks, 71–72, 73
 Journals
 online, 225
 submission of research to, 182, 308,
 309, 310, 315, 317, 319

K

Kendall Tau test, 290–291, 359
 Kruskal–Wallis test, 280, 359
 Kuder–Richardson 20 and 21, 130, 359

L

Laboratory research, 359
 Language background, participants',
 104–105, 109–110, 118, 126
 Language choice, 145–146
 Language related episodes (LREs), 128–129
 Learnability in UG, 48
 Learners, use of term, 25n
 Learning
 diaries facilitating, 177
 uptake sheets and, 202
 see also Instruction, effect of
 Legislation; U.S. National Research Act, 37
 Length of tests, 50–51, 62
 Libraries, 18
 Likert scale, 54, 55
 Limitations in research, 15, 19, 299,
 302–304, 314
 Line graphs (frequency polygons), 251, 256
 Linear regression, 289–290, 359
 Literature reviews, 5, 7–9, 19, 310–11

Literature searches, 18
 Location of study, 115, 118, 123
 Logistics, 206–209, 311
 Longitudinal studies, 111–114, 171, 303
 diaries, 204
 participant maturation, 115, 118
 participant mortality, 111–114, 118
 LREs (language related episodes), 128–129

M

Machines, transcription, 225, 367
 Magnitude estimation, 56–58, 359
 Mann–Whitney U test, 279, 280, 359
 MANOVA (multivariate analysis of vari-
 ance), 277, 360
 Map tasks, 71–72
 Materials, research, 138–141
 pilot testing, 43–44, 138, 141
 research reports and, 5–6, 9–11, 303
 Mean, statistical, 255–256, 262, 263, 307, 359
 and standard deviation, 259–261, 262
 limitations, 258–259
 see also SED; SEM
 Meaningfulness and significance, 267–268
 Median, 254, 262, 263, 307, 360
 Medical experimentation, 38–39, 244–245
 Memory, 56, 85
 Meta-analyses, 215, 283–284, 360
 Microphones, 66, 206–207
 Modality of tests, 51, 94, 96
 Mode, statistical, 254, 262, 263, 307, 360
 Mortality, participant, 111–114, 118, 149, 362
 Motivational characteristics, 145–146
 Moving window tasks, 63–64, 361
 MSB and MSW formulæ, 348
 Multimedia, 225
 Multiple choice formats, 59n
 Multiple methods research, 164, 170,
 181–182, 307, 360
 see also under Classroom research; Tri-
 angulation
 Multiple regression, 290
 Multivariate analysis of variance
 (MANOVA), 277, 360

N

Narratives, elicited, 87–89, 355
 National Commission for the Protection of
 Human Subjects of Biomedical

- and Behavioral Research, U.S., 37, 39
- National Institutes of Health, U.S., 26
- National Research Act, U.S., 37
- Naturalistic settings, 2, 86–87, 163, 165, 221–222, 361
- Negotiation vs. recasts, 46–47
- NIH (U.S. National Institutes of Health), 26
- Nominal data, 226–227, 278, 361
- Normal distribution, 261–263, 361
- Noticing, 3, 44, 102, 177, 182, 202
- Null subjects, 20, 157
- Nuremberg Code, 26, 38, 39
- O**
- Objectivity, 2, 188
- Observation, 37–38, 75, 76, 360
 - classroom, *see following entry*
 - participant, 170, 176
 - qualitative research, 165, 169, 175–176
 - self, 77
- Observations, classroom, 186–201, 351
 - in action research, 218
 - coding schemes, 190–201
 - COLT, 191–192, 193, 196–197, 200
 - custom-made, 201
 - TALOS, 190, 191–192, 193–195
 - consent for, 188–189
 - data reduction, 200
 - debriefing instructor following, 189–190
 - Hawthorne effect, 187–188
 - obtrusive observers, 187, 188–189
 - procedures and coding schemes, 190–201
 - triangulation, 201
- Observer's paradox, 29, 30, 176
- Obtrusive research(ers), 2, 187, 188–189
- OHRP (Office for Human Research Protections of DHHS), 26, 37
- OHSR (Office of Human Subjects Research of NIH), 26
- Omega², 282, 349, 361
- One-shot designs, 156–157, 158, 361
- Open role-play, 91, 360
- Open-ended processes, 93, 163, 164, 165, 170
- Operationalization, 105, 362
- OPI (Oral Proficiency Interview), 110
- Oral data, 66, 222–225
 - see also Recordings*
- Oral Proficiency Interview (OPI), 110
- Order of test items, varying of, 50, 114
- Ordinal scales, 227–228, 288, 362
- Organization of data, reporting of, 313–314
- Outliers, 254, 257–258, 362
- P**
- P-value, *see Probability*
- Parameters, 362
- Parents of participants, 34, 115, 209
- Partiality and impartiality, 163–164
- Participant observation, 170, 176
- Participants, 362
 - acknowledgment of, 317–318
 - characteristics, and validity of study, 109–111, 114–115, 118
 - debriefing, 30, 318
 - inattention, 114–115, 116, 118, 257–258
 - maturation, 115, 118, 149, 155
 - mortality (dropout rate), 111–114, 118, 149, 362
 - observer as, 170, 176
 - post-research feedback to, 318
 - qualitative research and point of view of, 167–168
 - research reports on, 5, 9, 114, 118, 312
 - researchers' own students, 34–35
 - terms for, 9n, 25n
 - voluntary, 34–35, 39, 40
 - see also Age; Anonymity; Attitudes; Compensation; Confidentiality; Fatigue; Gender; Language background; Protection of subjects; Sampling (participant selection)*
- Participation, researcher's, 170, 176
- Passives, Japanese, 45–46
- Past time/tense
 - English, 139, 171, 232–233
 - Spanish, 87–88
- Pearson product-moment correlation, 244, 286–290, 347, 362
- Pedagogy, implications of research for, 301–302, 304, 314
- Peer revision processes, 181–182
- Percentage agreement, simple, 243, 244
- Phenomenological tradition, 166
- Picture tasks
 - acceptability judgment, 52
 - description, 66–67

- elicited narrative, 89
 - spot the difference, 67–71, 140–141
 - Pilot studies, 41, 43–44, 362
 - coding, 248
 - consensus tasks, 74
 - elicited imitation, 56
 - interaction-based research, 65
 - requesting IRB permission to use data in main study, 44
 - materials testing, 43–44, 138, 141
 - spot the difference tasks, 71
 - stimulated recall and, 79
 - Planning
 - of task, 240–241
 - time, for response, 76, 88, 89
 - Populations, 362
 - See also* Samples
 - Post-hoc analysis, 363
 - Post-research concerns, 318
 - Posttests, 101, 363
 - delayed, 149, 354
 - equivalence with pretests, 116–117, 118, 130, 149
 - participant mortality and, 111
 - posttest only design, 149–150, 158, 363
 - pretest/posttest design, 148–149, 158, 363
 - Practice effects, 116, 118
 - Practitioner research, *see* Action research
 - Pragmatics, 171, 174
 - research, 47–48, 85–92
 - Predictive validity, 108, 363
 - Pretests, 116, 148–149, 158, 363
 - Posttest equivalence, 116–117, 118, 130, 149
 - Pretest/posttest design, 148–149, 158, 363
 - Previous work, relation of results to, 301, 304, 308
 - Probability, statistical, 264–268, 281, 363
 - of Type I and Type II errors, 266–267, 272, 368
 - Problem-solving and think-alouds, 79–85
 - Procedures, reporting of, 6, 11–12
 - Processing instruction, 213
 - Processing paradigm, 2, 61–64, 156, 163, 165
 - Processing time, discourse markers and, 76
 - see also* Reaction time
 - Proficiency levels, 110–111, 118, 120, 126
 - Pronouns, child acquisition of, 20–21
 - Proposals, research and grant, 319
 - Protections for human subjects, 39, 40
 - Protocols
 - research, 40–41, 364
 - verbal, in action research, 218
 - PsyScope computer software, 63n
 - Purpose, concealment of study's, 30–31, 50, 117, 119, 149, 176
 - Purpose sample, 122–123, 128, 363
- Q**
- Qualitative research, 2, 162–184, 363
 - characteristics and definitions, 2, 162–166, 363
 - in classroom settings, 212
 - data collection and, 165–166, 167–178
 - see also* Case studies; Diaries/journals; Ethnographies; Interviews; Observation
 - descriptions, 162, 306–307
 - descriptive research distinct from, 167
 - dynamism, 2, 169
 - emic perspective, 165, 168, 169, 170
 - group focus, 168
 - holism, 163, 165, 168, 171
 - hypotheses and questions, 164, 178–179, 182
 - informant checking, 165
 - multiple methods, 170
 - naturalistic settings, 2, 163, 165
 - and quantitative research, 2–5
 - combination methods, 2–3, 4, 164, 182, 307
 - differences between, 2, 5, 16, 93, 96, 165–166
 - questionnaires, 93, 96
 - researcher's participation in, 170
 - sociocultural context as focus, 168
 - statistics use in, 307
 - see also under* Coding; Contexts; Data analysis; Generalizability; Questions, research; Triangulation
 - Quantification, 363
 - Quantitative research
 - characteristics and definitions, 2–5, 363
 - see also* Qualitative research (and quantitative research) and individual items throughout index
 - Quasi-experimental research, 146, 363
 - Questionnaires, 75, 92–96, 364
 - in action research, 217–218
 - exit, and extra-experimental factors, 114–15, 118

- open and closed-ended items, 93
- Question formation
 - coding of data on, 234–236, 237
 - display questions, 169
 - recasts and development of, 234
 - spot the difference tasks and, 140–141
- Questions, research, 1, 16–23, 100, 365
 - and data collection method, 44–48
 - feasibility, 18–19
 - and hypotheses, 19–21
 - identification of, 16–23, 163, 302
 - open-ended, 164
 - in qualitative research, 16, 163, 164, 304
 - research reports on, 298, 300, 304
- Questions, test
 - distractor, 50
 - open-ended and closed-item, 93, 96
 - varying order of, 50, 114

R

- Race/ethnicity data, 126
- Random number generators, 121–122
- Random sampling, *see under* Sampling
- Range, 364
- Ranking procedures, 56–58, 227–228
- Ratio scale, 364
- Reaction time, 364
 - and data collection, 51, 56, 62–64
 - outliers, 257, 258
 - software, 246
 - see also* Processing time
- Recall
 - immediate, 85
 - see also* Stimulated recall
- Recasts, 3, 13–14, 46–47, 102, 234
- Recordings, audio or visual, 78, 175
 - anonymity issues, 28
 - automatically recorded data, 151
 - in classroom research, 206–209
 - equipment, 66, 206–209
 - video, 78, 91–92
 - see also* Transcription
- References, citation of, 6, 16, 315–316, 318
- Reflexives, 58–59
- Refugee populations, 28, 29
- Regression, statistical
 - linear: 289–290, 359
 - multiple, 290, 360
 - tests in SPSS package, 291
- Regression line, 364
- Relative clauses, 46–47, 107
- Reliability, 128–130, 364
 - acceptability judgments, 151
 - instrument, 129–130, 219, 312, 358
 - rater, *see under* Coding
 - transcription, 312
- Remuneration of participants, 34, 211, 318
- Repeated measures design, 150–151, 158, 364
- Repetition, task, 57–58
- Replication, 2, 21–23, 128, 230, 364
 - in qualitative research, 180
 - and generalizability, 22, 124
 - research reports and, 21n3, 303, 311
- Reporting
 - self-, 77
 - verbal, 77–85, 367
- Reports, research, 5–16, 297–321, 365
 - abstracts, 5, 7, 318, 310, 320, 350
 - acknowledgments, 314, 317–318
 - appendices, 6, 16
 - audience as consideration, 309–310
 - author note, 317–318
 - of case studies, 305–306
 - checklist for, 308
 - citations, *see references below*
 - conclusions, 6, 15, 299, 302–304, 314
 - constraints on, 308
 - discussion section, 6, 15, 298–302
 - final touches and formatting, 318–320
 - footnotes and endnotes, 316
 - front and back material, 318–319, 319–320
 - future research suggestions, 15, 299, 302–304, 314
 - gender-neutral writing, 319
 - introduction, 5, 7–9
 - in journals, 182, 308, 309, 310, 315, 317, 319
 - on limitations of study, 15, 299, 302–304, 314
 - literature reviews, 5, 7–9, 19, 310–11
 - on logistics, 311
 - methods sections, 5–6, 9–13, 304, 307
 - notes sections, 6, 15–16
 - on outlying data, 257–258
 - presentation, 314
 - on previous research, 301, 304, 308
 - on qualitative research, 170–171, 182, 304–307
 - on quantitative research, 298–304
 - on rater reliability, 244, 245–246
 - references, 6, 16, 315–316, 318

- on results, 6, 13–14, 298, 300, 304
 - combined with discussion, 298, 300, 304
 - possible reasons for, 300–301, 304
 - significance or implications, 301–302, 304, 314
- style guides, 315–316, 319–320
- tables, 317, 319
- on theoretical framework, 306, 308
- title page, 5, 6–7
- on transcription, 312–313
- see also under* Classroom research; Coding; Confidentiality; Data analysis; Data collection; Ethical issues; Generalizability; Graphic representation; Hypotheses; Materials; Participants; Questions, research; Replication; Statistics; Triangulation; Variables
- Representativeness, 123–124, 364
- Requests, child ESL learning, 171–172
- Research, 364
 - ethics, protocols, questions, reports, *see under operative word*
- Respect for subjects, 40
- Results, *see under* Reports, research
- Retroactive use of data, 44
- Review, *see* Institutional review
- Risks to subjects, 27, 30, 34, 37
- Ritualized uses of language, 169
- Role plays, 47–48, 91, 351, 362

- S**
- Sampling (data), 247, 354, 364
- Sampling (participant selection), 37, 364
 - and generalizability, 123–124
 - non-random, 122–123, 143–144, 146
 - convenience, 122, 128, 352
 - and counterbalancing design, 143–145
 - intact classes, 141–143, 150, 358
 - purposive, 122–123, 128, 363
 - systematic, 122, 366
 - random, 119–122, 128, 146, 364
 - cluster, 120–121
 - simple, 120, 365
 - stratified, 120, 128, 366
 - reporting of, 312
 - size of sample, 123–124
 - and validity, 119–123, 128
- SAS statistical package, 226, 291n, 365
- Scheffé test, 275, 365
- Scoring, 12, 54–55, 227–229
 - standard scores, 263–264, 366
 - see also* Interval scales; Ordinal scales
- Second language; use of term, 185n
- SED (standard error of the difference between sample means), 269–270, 346, 364
- Segmentation of data, 210–211, 247, 354
- SEM (standard error of the mean), 269, 347, 366
- Sentence interpretation, 61–62
- Sentence matching, 59–61, 365
- Setting for research, 115, 118, 123, 128
- Significance
 - and meaningfulness, 267–268
 - of results, 301–302, 304, 314
 - statistical, 267–268, 270
- Small group, 365
- SOC (suppliance in obligatory contexts), 232–233, 366
- Sociocultural research, 164, 168, 169–170
- Sociolinguistics research, 85–92
- Soft vs. hard data, 2
- Software
 - for psycholinguistic research, 63n
 - reaction time, 63n, 246
 - reference checking and formatting, 316
 - speech recognition, 225
 - statistical packages, 226, 280, 282, 291–292
- Sounding out words, 76
- Spearman Rank Correlation Coefficients
 - calculation, 244
- Spearman rho calculation, 290–291, 347, 365
- Speech recognition, automatic, 225
- Speed of response, 51, 257, 258
 - see also* Reaction time
- Split half procedure, 130, 365
- Split methods research, 164, 365
- Spot the difference tasks, 67–71, 140–141
- SPSS statistical package, 226, 291, 366
- SSB (sum of squares between) formula, 348
- SST formula, 348
- SSW (sum of squares within) formula, 348
- Standard deviation, 259–261, 262, 307, 366
 - pooled, 349
- Standard error, *see* SED; SEM
- Standard scores, 263–264, 366
- Statistics, 2–3, 4, 250–296, 347–349, 366

- computer packages, 226, 280, 282, 291–292
 critical value, 280–281
 descriptive, 250–261, 307
 see also Central tendency, Dispersion, *and* Frequency, measures of effect size, 282–283
 eta², 282, 348, 355
 factor analysis, 290–291
 formulæ, 347–349
 graphic representation, 251, 253, 252–254, 255–256, 285–287, 289, 307, 317
 and hypotheses, 100–101, 270–271, 281
 inferential, 269–280, 357
 prerequisites, 269–271
 see also Critical value; Freedom, degrees of; Hypotheses (one- and two-tailed); SEM; *and* non-parametric *and* parametric *below*
 meta-analyses, 283–284
 non-parametric, 271–272, 277–280, 361
 see also Chi square, Fisher's, Friedman, Kruksal-Wallis, Mann-Whitney U, *and* Wilcoxon Rank Sums tests
 normal distribution, 261–263
 parametric, 271–272, 272–277, 362
 see also ANCOVA; ANOVA; MANOVA; T-tests
 and reliability, 130
 and research reports, 314
 standard scores, 263–264
 strength of association, 137, 282, 366
 tables, 251, 252, 280–282, 317
 omega², 282, 349, 361
see also Correlation data analysis; Duncan's multiple range test; Probability; Scheffé test; Significance; Split half procedure; Tukey test
- Stimulated recall, 78–79, 366
 With classroom research, 201, 203
 for extra-experimental factors, 114–15, 118
 on noticing and interaction, 44
 and quantitative research, 307
- Story completion tasks, 72
 Story sequencing tasks, 72, 73
 Strategies-based research, 75–85
 Style guides, 315–316, 319–320
 Subject-verb order, 156
 Subjectivity, 2, 174, 204
 Subjects, as term for participants, 9n, 25n
- Sum of squares formulae, 348
 Suppliance in obligatory contexts, *see* SOC Surveys, 92–96, 167, 173, 366
 SYSTAT statistical package, 291n, 366
- ## T
- T, distribution of, 280–281
 T score, 347, 365
 T-test, 272–274, 279, 280, 348, 367
 matched, 362
 paired, 272, 273, 280, 348, 362
 in SPSS statistical package, 291
 T-units, 231–232, 368
 Tables, 13, 14, 319
 statistical, 251, 252, 280–282, 317
 TALOS (Target Language Observation Scheme), 190, 191, 193–195
- Tasks
 closed, 65
 one and two-way, 65, 71–72
 open, 65
 planning, 240–241
see also picture tasks; spot the difference; jigsaw; consensus; consciousness-raising
- Teachers
 collaboration with researchers, 219–220
 diaries, 203–205
 post-research feedback to, 318
 principles and classroom practices, 94, 95
 research by, *see* Action research
- Tense-aspect morphology, 232–233
 Test-retest method, 129
 Test-taking abilities as variable, 104
- Theory
 implications of study results, 301–302, 304, 314
 study framework, 44–48, 306, 308
 quantitative and qualitative approaches, 166
- Think-alouds, 77, 79–85, 218, 367
- Time
 timing of tests, 51, 78, 85
 time-series designs, 152–155, 158, 188, 367
see Planning (time); Processing time; Reaction time; *and under* Coding
- Transcription, 222–225
 broad and narrow, 222

conventions, 222–224, 313, 342–344, 345–346, 367
 machines, 225
 research reports on, 312–313
 technology and, 206, 225, 246, 367
 Transferability, 180, 367
 Transition anxiety, 177
 Travel, biodata on, 126, 127
 Triangulation, 368
 action research, 217
 classroom research, 201, 206, 217, 306
 investigator, 359
 methodological, 360
 and motivational information, 176
 multiple methods, 170, 181–182
 observations, 176, 201
 qualitative research, 165, 169, 170, 181–182
 research reports and, 306, 307
 theoretical, 367
 Truth-value judgments, 46, 58–59, 368
 Tukey test, 275, 368
 Tuskegee Syphilis Study, 38–39
 Type I and II errors, 266–267, 272, 368
 Types of research, 2–5
 Type-token ratio, 369

U

UG, *see* Universal Grammar
 Universal Grammar, 48–61, 156–157
 Uptake sheets, 201–203, 369

V

Validity, 106–128, 369
 action research, 219
 construct, 107–108, 352
 content, 107
 criterion-related, 108, 353
 external, 119–128, 179n, 356
 face, 107, 143, 356
 internal, 109–119, 358

predictive, 108, 363
 qualitative research equivalents, 179n
 reporting of, 312
 VARBRUL statistical package, 291–292, 369
 Variables, 101–105, 369
 in classroom research, 215
 control, 104–105, 352
 dependent, 102, 103, 354
 independent, 102, 103, 137–138, 357
 individual emotional, 177
 intervening, 104, 215, 358
 measurement scales, 105–106, 313–314
 moderator, 103–104, 151, 360
 operationalizations, 105
 research reports on, 309, 311, 313–314
 Variance, statistical, 259–260
 see also ANCOVA; ANOVA; MANOVA
 Verb inflection, child acquisition, 20–21
 Verbal reporting, 77–85, 369
 Verification, 2, 22
 Video recordings, 78, 91–92
 Vocabulary studies, frequency of words in, 150
 Voluntary participation, 39, 40, 34–35

W

Web, Worldwide, 18, 29, 97
 Wilcoxon Rank Sums test, 279, 280, 369
 Window, moving, 63–64, 360
 Within-group design *see* Repeated measures design
 Words
 count/sentence, and topic, 151–152, 153
 focus on individual, 76
 frequency indexes, 116
 Writing studies, 116, 164, 182, 238–240

Z

Z scores, 347, 369