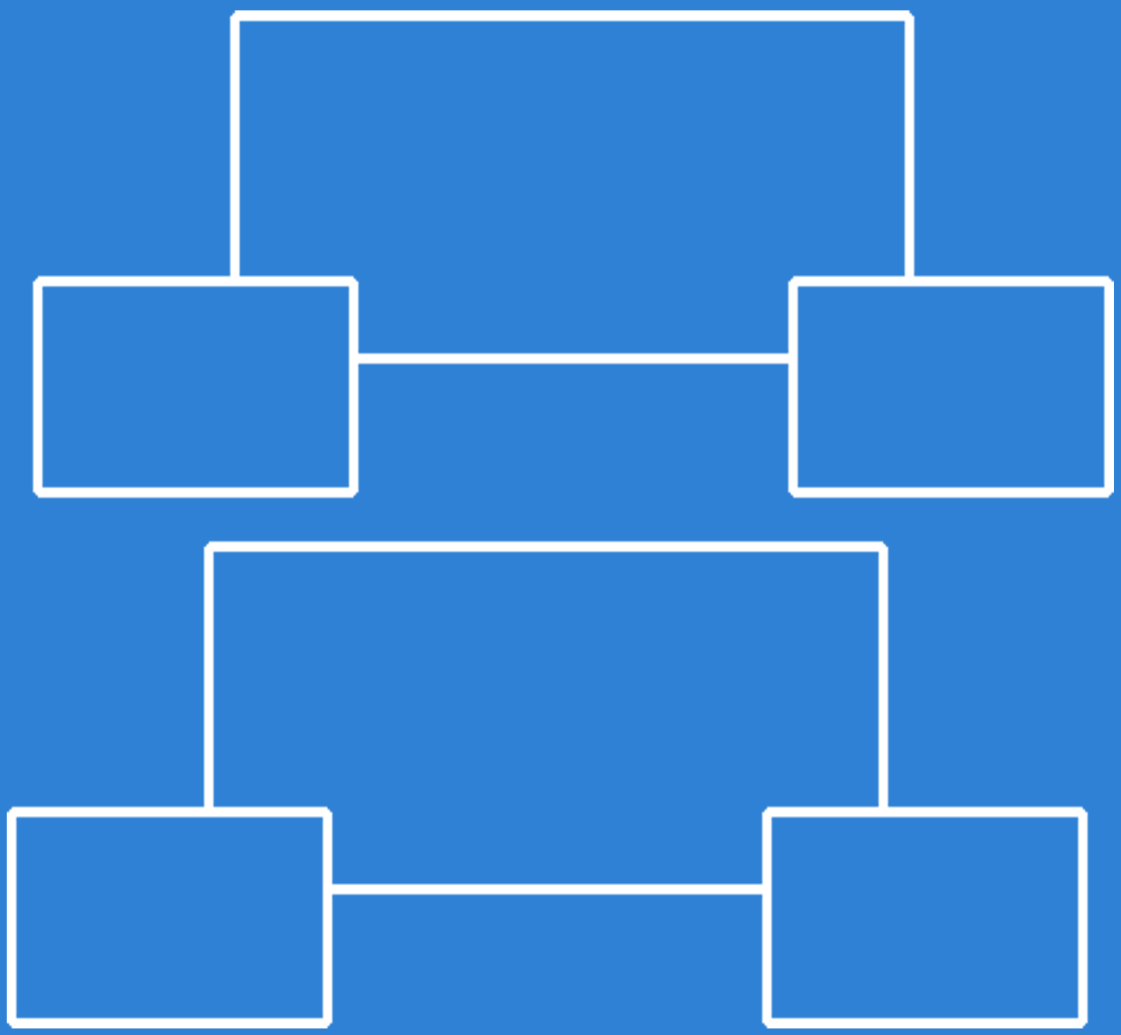
An Introduction to Second Language Research Methods

Design and Data



Dale T. Griffee

TESL-EJ Publications

tesl-ej.org

eBook edition 2012

Produced in the United States of America.

First Edition

Library of Congress Cataloging-in-Publication Data

An Introduction to Second Language Research Methods: Design and Data / by Dale T. Griffee
p. 21 x 28 cm. 213 pp.
Includes bibliographical references
ISBN 10: 0-9823724-1-8 ISBN 13: 978-0-9823724-1-8 (eBook)
1. Research Methods 2. Education 3. Language and Languages—Study and Teaching

About the Author

Dale T. Griffee directs the International Teaching Assistant (ITA) program at Texas Tech University where he teaches ITAs, academic writing, research methods and program evaluation. He holds an MAT in ESL from The School for International Training and an Ed. D. from Temple University, Japan. He was the series editor for the JALT (Japan Association of Language Teachers) Applied Materials series, and with David Nunan edited Classroom Teachers and Classroom Research, (1997) a compilation of articles on classroom research. His most recent publication is as fourth author of *English Communication for International Teaching Assistants* (Waveland Press, 2010). He welcomes questions and comments at dale.griffee@ttu.edu.

This book is set in Cambria font.

Editor: M.E. Sokolik, University of California, Berkeley

An Introduction to Second Language Research Methods: Design and Data

Dale T. Griffee

TESL-EJ Publications

Berkeley, California, USA

An Introduction to Second Language Research Methods:

Design and Data

PART ONE
Getting Started

MY APPROACH TO RESEARCH

First of all, congratulations! You graduated from various schools and now you are interested in research. At some point you decided to be a language teacher. I'm an ESL teacher myself, but whatever language you teach, congratulations again and welcome to the club.

In college, especially as undergraduates, we learn knowledge or knowing, but we also learn action or doing. For example, in drama departments, students learn the history of the theater (knowing), but they also learn how to put on a play (doing). In second language teaching there is the same relationship between knowing things and doing things (Dunne & Pendlebury, 2003). Our knowing includes how language is processed and how it is acquired; our doing includes teaching and answering questions.

My approach to research is that we should not concentrate on doing things to the point that we forget that doing is ultimately based on knowing. Always thinking about doing at the expense of knowing blinds us to the relationship between what we know (our theory) and what we do (our practice). Teachers usually get it half right and half wrong. We get it that knowing without doing is pointless, but we don't always get it that doing without knowing is blind.

All teachers know that a classroom is an active place. In fact, many teachers and students are often doing multiple things at the same time. Our brains are constantly engaged, and it is hard to stop, record our actions, and reflect on them in the midst of teaching. Yet despite the buzz of activity in a classroom, all our actions are done for a purpose, and the purpose contains within it a reason. This implies that every action we do has a rationale--a theory that guides the action by providing a motivation to the actor for the action. What we do is dependent on what we know. Teachers often call this "what works for me" because we know what *we* think, but we do not know how or even *if* it applies to other teachers. How does this practical stance come about? Where did we learn it?

The answer is our background and experience. I watched my teachers and tried to remember what they did. I also learned to teach by reading and listening to the advice of others, but most of all I learned by direct and usually painful experience in the classroom. So, for me and maybe most teachers, action is supreme and tends to crowd out knowing. This is not necessarily a bad thing. Action requests of knowing that it has a connection to teaching, and that the connection between them be made explicit. But, if it is true that what we do is, in the final analysis, based on what we know, then we ignore knowing at our peril. I think teachers are correct in insisting that knowledge (ideas, principles, theory) be relevant to doing (teaching). The trick I want to learn is how to articulate my own theory and how to integrate that theory with those of others. And one way to do this is through research.

I am on a journey and you are, too. As undergraduates, we were students who took knowledge acquisition as our job. We studied, sometimes, what we were told to study by authorities (college and university teachers), and we usually believed that what we studied was true. We studied

vocabulary, which contained other people's definitions and other people's theories. Now we are participating in the construction of that knowledge. For me, two of the shocks of becoming a graduate student were the realization that each definition I learned was actually a low level articulation of a theory, and second, in many cases there were fights over the meaning of those words (because definitions are actually competing theories). In other words, we are in a world in which people fight over words and their meanings, and knowledge is created, not taken for granted. I once heard a story about three baseball umpires that illustrates this point. They were being interviewed by a newspaper reporter who wanted to know how they called balls and strikes. What the reporter didn't know was that one of the umpires operated out of a classic philosophy, one had a modern theory, and the third had a postmodern theory. The reporter asked, "How do you know the difference between a ball and a strike?" The classical umpire said, "I call them as they are." The modern umpire said, "I call them as I see them." The postmodern umpire said, "They aren't anything until I call them." True enough, but what all three umpires don't fully appreciate is that all of them had theories, which filter their beliefs and determines their actions.

My approach to research is that it is not enough for me to know my way around a classroom. I want to become aware of what I believe and why I believe it; I want to be able to create and construct my knowing, not (only) so I can become a more accomplished knower, but so I can be in charge of my doing which is teaching. That's what I think research is all about.

Reference

Dunne, J., & Pendlebury, S. (2003). Practical reason. In N. Blake, P. Smeyers, R. Smith, & P. Standish (Eds.), *The Blackwell guide to the philosophy of education* (pp. 194-211). Malden, MA: Blackwell.

CHAPTER ONE

HOW TO GET STARTED

The major argument for educational research, carried out by people who are closely involved with teaching, is that teaching is a complex activity, and no one else will produce the kind of research needed. (Brumfit, 1995, p. 36)

In this chapter you will learn: What TREES are, some of my thoughts and assumptions about research, some of your thoughts and assumptions about research, and how to get started by thinking of a problem to research.

Introduction

In this chapter, we will explore some ideas on how to get started in your research project. As you read, pay attention to questions that come to you. At the end of this chapter you will see a large box. In that box write a question or two that popped into your mind as you read this chapter.

What does TREE mean?

TREE stands for Teacher-Researcher-Educator-Evaluator. I use this term because it best describes me as a classroom teacher and a researcher. Sometimes I think of myself as a teacher who is also a researcher, and sometimes I think of myself as a researcher who is also a teacher (Stewart, 2006). I am also an educator and an evaluator; as an educator I am interested in classroom issues, as an evaluator I want to know what works in my classes. Thus, there is a relationship among being a teacher, a researcher, an educator, and an evaluator. When I write about research, I sometimes refer to myself and to my readers in one of those roles and sometimes in another; this is confusing because no matter which term I use, I want to include the others. Therefore, I coined the term TREE to include all of these roles.

Some of my assumptions about research

1. Most teachers see themselves primarily as classroom instructors, and secondarily as researchers, if at all. One reason for this assumption is that most of us got started in teaching because we found that we loved the interaction of teaching, with little or no thought of research. Another reason is that we often have not been trained to be researchers, and as a result the thought of doing research often annoys us or scares us.

2. Teachers should also be researchers because I agree with Brumfit (1995) who says that teachers need to do their own research because nobody is going to do our job for us.

3. One way to begin thinking about research is to study the structure of the published research papers. Although no two research papers are exactly alike, there is a common structure which identifies the genre, and it can be helpful to be familiar with that structure.

4. In addition, dividing research into quantitative data (numbers and statistical analysis) and qualitative data (verbal or narrative reporting) is not helpful because it shunts us off into one area or the other, and such thinking doesn't encourage us to use all of the tools and ideas we otherwise could.

5. Research and evaluation are activities that every teacher engages in, except that we do not usually call it research. Whenever we walk into a classroom, we are alert to all kinds of input that we reflect on, categorize, and learn from. In that sense, research is what happens when we do our normal job.

6. Ideas for research come from many sources, such as thinking about our teaching, going to conferences, talking and listening to others, reading books and journals, becoming aware of a problem, and taking courses, especially graduate level courses.

7. If we are to survive and succeed as individuals as well as professionals, we have to attend conferences and give presentations. No matter the size of our school or the city in which we are located, we can be isolated; we need to constantly be reaching out and networking. Attending conferences is one way to do that.

8. Everyone feels inadequate. Remember the first time your first bicycle ride, your first kiss--in fact, the first time you did anything? Getting started is not a pretty sight. When starting research, we teachers are in the same position that our students are in most of the time, trying to get started while feeling inadequate.

9. Finally, anyone can do research. If you have enough knowledge and skill to be a teacher, you have enough knowledge and skill to be a researcher. The question is, why would anybody want to? I have two answers. First, if we don't do research to answer our questions, nobody else will do it for us. Second, if we don't engage in research, we will be doomed to keep repeating our experiences.

What are some of your ideas about research?

In order to begin to answer this question, try answering the following questionnaire, which I call the Preferences Data Questionnaire, or PDQ. There are only eight questions. Your answers may give you an insight into the type of research designs you are attracted to. To take the PDQ, read the eight statements. As you read each of the items, circle the number that best fits your feeling about the statement. If you aren't sure, guess.

Preferences Design Questionnaire (PDQ)

1. I like to observe personally what I am researching to get data.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

2. Interviewing people is a good way to do research.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

3. Questionnaires show what people really think.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

4. Letting people mark or write answers to questions is a good way to collect information.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

5. Tests can show what students have learned.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

6. If you want to know something, ask questions and listen to what people say.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

7. If I am doing research on my students, I like to watch them carefully and notice what they do.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

8. Although not perfect, test scores can be a good indicator of learning.

No	Not Sure	Maybe	Probably	Absolutely
1	2	3	4	5

How to score the PDQ questionnaire

Total your scores for items 3, 4, 5, and 8 and record it under A

Total your scores for items 1, 2, 6 and 7 and record it under B

A	B
Your score on item 3 _____	Your score on item 1 _____
Your score on item 4 _____	Your score on item 2 _____
Your score on item 5 _____	Your score on item 6 _____
Your score on item 8 _____	Your score on item 7 _____
Total _____	_____

If your score in the A column is higher than your score in the B column, it may indicate you prefer looking at research analytically, and you might like to use tests and analyze them using statistics. You might like the experimental research design. You also prefer questionnaires that provide numbers in much the same way that the one you just completed did. In that case, you might like the survey research design.

If your score in the B column is higher than your score in the A column, it may indicate you prefer considering the whole picture and not breaking things down into parts. You might be drawn to observation and informal situations. Perhaps you like to listen to what people have to say. In that case, you might be interested in a research design such as case study or action research.

What do your PDQ scores mean?

First, work with a partner or in a small group**.** Tell your partner what your PDQ A and B scores are. Listen as they tell you their scores.

Second, in your opinion, are your column A scores very different (more than three points apart) or practically the same (three or fewer points apart) as your column B scores?

Third, what is your best guess as to what your PDQ scores might say about you and how you might begin a research project?

How I usually get started

There may not be any one best way to start a research project that works for all of us all the time. However, as I reflect on my own experience, I seem to have a general starting point. I often start from a problem, and from the problem I work to formulate a research question or questions I want to answer.

Where do research questions come from?

Formulating a problem may not be easy. This could be because as readers of research, we often see the results of another person's research, but we do not know the process the researcher went through to get the idea for it. I decided to ask faculty, graduate students, and undergraduate students how they get ideas for research projects. After informally interviewing these groups of researchers, three sources for research ideas emerged: research ideas that originate outside ourselves, ideas that come from within ourselves, and ideas that come from current research in our field of interest.

Research ideas that originate outside ourselves

More experienced researchers (MERs), such as teachers, state that they are receptive to questions from their students. For example, teachers can get ideas for research from the questions that students ask. Additionally, MERs are likely to attend professional conferences, which they report as another way of getting ideas. They attend sessions on topics of interest, and come away with handouts, ideas, and in some cases, the presenter as a research collaborator. Finally, experienced researchers often look for or are approached by other persons who suggest topics and ask them to partner with them on research projects.

Less experienced researchers (LERs) tend to listen to their professors. They listen carefully to the course instructor as he/she mentions research ideas in class or in meetings after class. In graduate school, I used to take notes, especially during the first class, when my professors gave their opening lecture. Often they would give a review of the subject with possibilities for research projects. I could nearly always find a topic. Sometimes LERs ask their professors directly for research ideas, and sometimes they are able to work with professors and joint projects. Attending conference presentations can also give LERs ideas. Finally, less experienced researchers have told me that they talk to their friends both in and out of class for research ideas.

Research ideas that originate from within ourselves

More experienced researchers have research experience that provides ideas and gives them a base from which to launch new research. MERs read and reflect on current research that provides them ideas. MERs pay attention to what they are interested in, and find the research aspect of that interest.

Less experienced researchers (LER) report that they think about what they want to know, and sometimes research projects come from an area or person they want to know more about. Their biggest challenge is to find a research area that they would find interesting enough to sustain them through the research process.

Research ideas that originate from research in the field

More experienced researchers are likely to be currently engaged in research; they know that whatever project they are doing, it is likely that it will spawn additional ideas for investigation. In other words, they see a research project as an area or multiple ideas, not an isolated idea that once investigated is complete. In addition, MERs read the literature in their area and as a result can identify gaps in research. They know that they can research those gaps. Finally, most textbooks are written by the very instructors that teach the courses in which those textbooks are used. When a textbook becomes outdated or when they have a new approach, MERs can nominate themselves to work on a textbook, which in many cases involves research.

Less experienced researchers report that they read generally and more specifically in the field they want to research. More general reading includes: reading newspapers, going online to search for a topic, and pleasure reading to get ideas. Specific reading includes: reading journal articles from previous research and the assigned text.

If you are new to research, how can you get a research idea?

1. Listen to your instructors in class as they discuss research areas. Make an appointment with your instructor, and then follow up. Give your instructor some idea of what interests you; don't expect to just be given a research idea. Your instructors probably had to work for their ideas, and they expect you to do the same.

2. Attending conferences on a topic area is always a good idea. There may be small conferences held nearby, especially those in your university and even in your department. A conference of any size is a marketplace of ideas waiting to be picked up free of charge.

3. Ask your librarian what journals and periodicals are available in your library; browse through them to find interesting topics. If you locate an article describing research that interests you, there is no reason you can't replicate the research. Get a copy of the article, study it, and see if you can do the same or similar research.

4. Think about an idea you are familiar with and would like to investigate further. Ask your instructors what they think about the idea. Get into the habit of talking to your fellow students as professionals about research.

5. Consider any assigned textbook as a source for research ideas. Look through the tables of contents for interesting chapters, and then skim and scan for ideas.

6. Another approach to getting started in research is Barkhuizen (2009, p. 116) who asks teachers new to research a series of seven open-ended questions. For example, his first question is, "I remember once in my classroom I had a very difficult time trying to …." All the questions are related, and together the answers might form a helpful narrative to get started.

DISCUSSION QUESTIONS

Task 1. Pick at least one of *my* assumptions (pages 9-10) and say why you agree or disagree with it.

_____

_____

_____

_____

_____

Task 2. List two of *your* assumptions about research. Think of an assumption as any thought, idea, or belief you have.

Assumption one _____

_____

_____

_____

Assumption two _____

_____

_____

_____

Task 3. Fill out the research proposal form. If some of the categories are not clear, write "undecided" and fill them in later. Assume you can change your mind at any time, but if a proposal or a complete research paper is required, also assume that time is limited.

Research Proposal Form

Date _____

Where did the idea for this research come from? _____

_____

_____

What is the purpose of this research? _____

_____

_____

State your research questions (RQs) _____

_____

_____

_____

For each RQ, state the purpose. (Why are you asking it, what do you hope to find?) For example, RQ1 will tell me X, RQ2 will tell me Y.

_____

_____

_____

_____

What data collection instrument could you use to gather data? _____

_____

_____

Who are your likely participants? _____

_____

_____

References for How to Get Started

Barkhuizen, G. (2009). Topics, aims, and constraints in English teacher research: A Chinese case study. *TESOL Quarterly, 43*(1), 113-125.

Brumfit, C. (1995). Teacher professionalism and research. In G. Cook & B Seidlhofer (Eds.). *Principle & Practice in Applied Linguistics: Studies in honour of H. G. Widdowson* (pp. 27-41). Oxford: Oxford University Press.

Stewart, T. (2006). Teacher-Researcher collaboration or teachers' research? *TESOL Quarterly, 40*, 421-430.

CHAPTER TWO

STRUCTURE OF A RESEARCH PAPER

Although research findings are, to some extent, always inconclusive, practices unsupported by research are even riskier. (Swaffar & Bacon, 1993, p. 143)

In this chapter you will learn: The format of the standard research paper, and how to structure, search, and write a research paper that includes a literature review. Even though every research paper is unique, there is an organizational pattern that many research papers tend to follow; I will refer to this as the standard form. Later, I will discuss possible alternative forms associated with qualitative research procedures. In this context, it is important to remember that the process of writing and research is recursive--you may write a section, move on to another section, and then go back and revise. In that sense, writing a research paper is not always a straightforward process. The standard form of the research paper can be used as a checklist at the early stage of your research project, and again later as a writing model at the end stage of your research. This is important because if one of the sections described here is missing from your paper, you may be less likely to convince readers that your research is sound.

Levels of headings

Three levels of heading are usually sufficient for most research papers, but see the American Psychological Association's (APA) format guide (2010, p. 62) for examples using up to five levels. Following is an example of a three-level heading. The first level is centered with title case* headings, the second level is flush left, italicized with title case side headings, and the third level is indented, italicized, lowercase paragraph headings ending with a period. Figure 1 shows how three levels look in terms of visual placement.

The Title

Creating a good title for a research paper does not usually receive attention, but for many TREEs it can be a challenge. The title of a research paper is important because it creates the first impression of a paper. People may decide to read or not read a paper simply by the title. If you send your manuscript to a journal, your title may determine to whom the journal editor assigns it for review. The *APA Publication Manual* recommends titles be 10 to 12 words long (APA, 2010, p. 23). A major consideration is that the title will be catalogued in various databases. Remembering that, create a title that describes your topic using keywords that can be used by

---

** Title case refers to a system in which the first letter of main words in a title are capitalized, but all other letters are in lower case. An example of title case is: War and Peace. You'll notice that the word 'and' is not capitalized.*

Title

Abstract

Method

Participants (begin on next line, indent and begin with a capital letter)

All students were … . (if you need subcategories, see next line)

First year students. First year students consisted of ...
(if two subcategories are needed, indent and begin the next line)

Second year students. Five second year students were from China … .

Materials

Procedures

Analysis (A proposal stops here, but includes references and appendixes.)

Results

Discussion

References

Appendixes

Figure 1. Three levels of headings in a typical research paper format

others in a search. For example, if you write a paper on basic writing, you would want others interested in basic writing to be able to locate your paper. Using a title which captures your feelings such as "The good, the bad, and the ugly," but which does not contain key search words can result in momentary satisfaction because this poetic title captures your feelings, but long-term loss may result because readers may not recognize what your topic is. This title could be changed to "Basic writing for ESL students: The good, the bad, and the ugly."

One strategy is to insert the term "working title" in front of your title in order to hold your initial thoughts and also to remind you that it can be changed. Another strategy is to list keywords from the paper, and using those key words, arrange them in multiple ways to create several possible

titles. Then describe your research to your colleagues while asking them to vote on the most appropriate title. A majority vote by your colleagues may point to the most appropriate title.

The Abstract

Abstracts force us to synthesize what our research is about; they are written primarily for three reasons: they are submitted as part of paper manuscripts to journals; they are sent as part of the application package for conference presentations; and they function as a summary that readers use to decide if they want to read the whole paper. An abstract must include a lot of information in a short space, and are typically limited to a specified number of words. Often conference application forms provide specific formats for abstract submission.

Below are some guidelines from Brown (1988) that might be helpful in writing an abstract. An abstract can include a statement of topic, the purpose of the article, a description of participants, a list of materials used in the research, an explanation of materials, the statistical analyses used, a summary of results, and implications for the field. Tuckman (1999) offers a slightly different list of suggestions. He suggests an abstract should be 100 to 175 words, single-spaced in block form, which means no indentations, and include sections on problem, method, results, participants, type of research design, statistical significance levels, and conclusions. Finally, use only standard abbreviations and acronyms in your abstract.

The Introduction

The introduction is different from most other sections of a paper in that generally, the word "Introduction" is not used. One just begins. Introductions to academic papers come in many sizes; there is no "one size fits all," but they must make sense to the reader. One way to grasp the structure of an introduction would be to read articles in one or two of your favorite journals to see how they do it.

Swales (1994) suggests beginning your paper by stating in a general way why this topic is interesting to your academic field. Second, include a statement of the problem using the present tense. Citation is optional for this step. State why this problem is important. This step has several possible versions. If you see a gap in the literature, your paper fills that gap. You can begin with words such as "however," or "nevertheless." If you are raising a question, your paper answers this question, if your are continuing a tradition, your paper is making a contribution, and if you are refuting a claim, your paper substantiates a counter claim. Many research papers put their purpose next followed by a literature review and research questions (RQs). Figure two shows the structure on the left and an example of the structure on the right.

A literature review can be understood as its own form of research, which is called *secondary research* and is sometimes referred to as *library research*. If you don't do a literature review, in other words, if you don't read any background information, then your only resource is your personal experience. It appears you are unaware of the published findings of others. The literature review keeps TREEs from "reinventing the wheel" in that it tells them and their readers what researchers in the field have done, so they can build on their results and not keep repeating

First, state why this problem is interesting to your field.	It is generally acknowledged that textbooks play an important role in language classes. For example, Author (19xx) states that textbooks provide significant language input.
Second, state a problem, contradiction, gap, or question using words such as: *however, nevertheless, yet, but.*	Nevertheless, some researchers (Author, 19xx; Author, 20xx) have shown that many text authors employ artificial dialogues that lack normal language features.
Third, state the purpose of the paper.	The purpose of this paper is to investigate current textbooks comparing their dialogues with those of native speakers.

Figure 2. One possible structure of an introduction

findings that may be irrelevant. If many or most practitioners in a field did not publish articles with literature reviews, it would become difficult for that field to progress because knowledge would not accumulate.

For years, it was a perplexing issue for me as to which tense to use in writing various parts of a research paper. It made sense to use the past tense in all sections because the research had already been accomplished. However, Swales and Feak (2004) have surveyed research papers using the IMRD (Introduction, Methods, Results, Discussion) model, and conclude that the present tense is more common in the introduction and discussion while the past tense is more common in the method and results section. While "common" does not mean "always", this is good advice.

How to write a Literature Review

The literature review is the star of the introduction. In fact, the literature review is so prominent that many TREEs, if they do not read journal articles carefully, may overlook other important parts of the introduction such as statement of the problem, the purpose, and the research questions.

The most obvious feature of a literature review is the discussion and citation of relevant published material, including journal articles, books, reviews, reports, conference papers, and even personal communication, all of which taken together can be called *the literature*. Citation of claims is important because at this point, readers are not interested in opinions. A literature review can be from one paragraph to several pages, but regardless of length, it should be a synthesis of what is relevant, not just a list of one thing after another, and it must directly support your research questions.

For many TREEs, there is a misunderstanding about what goes into a literature review and why. It is often believed that a literature review is a gathering and analysis of all the material that has been written on a topic or at least all the material that is relevant to your topic. This is obviously impossible. It may be helpful to look at your research question (RQ) The RQs guide the construction of a literature review. If material is not relevant to your research question, it is not relevant to your literature review. The literature review is also the location of theory in your paper. Even if there is no formal theory about your topic, the selection, argument, and synthesis of relevant papers constitute the working theory of your paper.

Purposes of a literature review

The literature review serves many purposes:

1. It motivates the study and provides background (Bill VanPatten, personal correspondence, February 14, 2009). In fact, in addition to the term literature review, terms such as background and context are becoming more common. There is a strong connection between the literature review and research questions. This connection is so strong that we should be able to read the literature review and almost anticipate the research questions.

2. It educates readers on your topic. Many literature reviews are, in fact, short histories of a topic.

3. A literature review identifies your intellectual history by telling readers what you have read and what traditions you draw from.

4. It identifies your sources. The literature review provides a paper trail indicating where to find the articles, books, and other resources used. In that sense, it is a source file for others to access.

5. It provides researcher and readers alike ideas for further research. Literature read for the literature review may give a researcher ideas as well as possible research designs to draw from, replicate, or modify. For example, in doing a literature review, you may come across problems mentioned by other researchers that you might not otherwise been aware of, and these problems can be addressed and/or can become research questions.

The downside of a literature review is that it is difficult to write. This is because it is not always easy to know what to include or exclude, sometimes it is difficult to find sources, and other times there is too much material. Knowing how to synthesize material is a major problem for most writers. Lastly, it is often difficult to know how or where to begin a literature review because it requires special search skills.

Two scenarios of getting started

The case of inexperienced Ralph

Ralph is a senior taking a graduate level research course for the first time, and a research paer including a literature review is required. Ralph has never written a research paper before and

certainly not a literature review. In fact, he really isn't sure what a literature search is or how to go about doing one. After much thought, Ralph collects himself and thinks, "I can do this." The problem is he is not sure what "do" and "this" really mean.

Ralph decided that he has two choices: he could talk to his professor, or he could visit the library. Since his professor had suggested in class that anybody having trouble with the paper make an appointment to talk about it, Ralph decides to try this first. He explains to his professor that he has no clear idea of a research topic, no idea of how to collect data, and no idea of what a literature search is or how to do it. The professor knows that Ralph is an undergraduate in a graduate course and probably has little specific content knowledge or search skills. He also knows Ralph, unlike some of the graduate students, is neither a teaching assistant nor teaching a class of his own. The professor suggests that Ralph use their class for his study, and that the topic be a survey of what first-year graduate students think about research. He tells Ralph to go to the library and locate at least five general reference books on research methods since these books often begin with a "What is research?" chapter. From this literature, Ralph is to synthesize his findings and make a model to answer the question, "What are three, four, or five aspects of research?" From each of these aspects, Ralph is to formulate one or two questions that reflect the aspect. He then can interview most or all of the class to ascertain what the graduate students think of each aspect, and write the report based on that data. Ralph goes to the library and asks the librarian how to locate these books. He locates five such texts, checks them out of the library, takes them home and begins to read and take notes.

The case of experienced Wanda

Wanda is an Assistant Professor who must do research and publish her work in order to keep her job and get tenure. Her experience includes a masters degree followed by three years of doctoral level courses in which empirical research papers were regularly assigned. She took an additional four years to research and write her doctoral dissertation. Then followed two years of working at another college, during which time Wanda did not do much research, perhaps as a natural response to years of sustained pressure working to complete her doctorate. Wanda is ready to undertake research again, but she is interested in an area outside her doctoral work. This means that the content is new, but she can draw on her previous research skills and experience.

Wanda first wonders what journals she can search. She brainstorms a list of journals she is familiar with that might have articles on her new topic. She wants to know which of these journals the library has and which it does not, so she goes to the library's web page and finds out. Those journals the library has in the stacks can be physically examined. Those journals the library does not have can still be searched online or in databases for their tables of contents. Wanda is interested in two things: journals that she can get her article published in and also journals that contain relevant articles for her literature review. Wanda searches the journals published in the last six years and copies relevant articles to read. From the on-line journal table of contents, she makes a list of possible articles and goes on-line to order or read the articles. She now has approximately nine articles, one of which has an extensive reference section which Wanda reads carefully; she is delighted to find at least an additional ten highly relevant citations, some of them in journals outside her field and previously unknown to her.

How do I know what to look for?

It is helpful to think of literature as a plural noun because multiple literatures might be involved. When beginning a literature review on international teaching assistants (ITAs), I decided to look for literature in the areas of ITA issues and performance test criteria. As I searched, found, and read articles, I modified my literature search categories to ITA issues, ITA performance test categories, performance test theory and construction, and test validation. Having these four literature categories helped me make my search more comprehensive.

How do I take notes?

The main problem with taking notes is that the notes tend to follow the content of the article or chapter being noted. Notetaking also tends to be influenced by whatever problem is currently on our mind. For example, when one article is about apples and oranges, if the TREE is thinking about apples, the notes tend to be about apples with little attention paid to oranges. On the other hand, if another article is mainly about oranges, the notes tend to be about oranges. This makes it hard to compare and contrast notes in order to arrive at a synthesis. Another problem encountered by TREEs in their notetaking is that some papers are long and complicated, leaving the TREE to wonder how much data should be included.

What usually happens

What usually happens is that we receive an assignment to do a research paper, go on-line or to the library, find some material, read the materials, take random notes, and from our notes write our literature review. (Random notetaking is taking notes on one paper (P1 in Figure 3)). Often we are sensitive to certain issues that are on our mind at the time. Then, later, we read another paper (P2), with another set of interests and issues. This random notetaking process makes it very difficult to write a synthesis, because our notes reflect our various interests at different times. As a result we often write a "beads-on-a-string" literature review-- a one-thing-after-another literature review. Author A says this, author B says this, and so on.

What's wrong with a "beads-on-a-string" literature review?

The problem is that not only is it boring to read, it is almost impossible to understand. It has no point, no structure, no sense of direction, and ultimately no meaning. After about a page or so of reading a beads-on-a-string, one-thing-after-another literature review, a reader begins to ask, "Why am I reading this?" and often, stops reading.

Literature Area			
	P1	P2	P3 etc.
Note contents	Apples and Oranges	Apples and Pears	Pears and Peaches
Literature Review	One thing after another beads on a string review, or a confusing mix.		

Figure 3. The flow of a beads-on-a-string literature review

What is lacking is a filter or a screen that could be used to unify the data from the papers to help write a synthesis. For example, categories from Boote and Beile (2005) can be adapted as seen in Figure 4. Asking and answering the same questions of all articles increases the possibility of achieving a synthesis. It is not important that you use these categories, but it is probably necessary for you to devise a systematic way of notetaking. It is also important that notetaking categories be related to purpose and research questions.

Some definitions of the terms in Figure Four

Coverage means the criteria for inclusion or which academic areas, here called literatures, to include in the literature review. *Synthesis* refers to a summary of various insights from various papers. Synthesis is what is generally accepted as traditional notetaking procedures. The three questions in the synthesis section are difficult to answer, but they push your depth of reading and understanding. *Methodology* refers to the design used in the paper. *Significance* means importance.

A literature review is not just a listing of points and issues made in one article after another, but a creative synthesis. A literature review provides readers with a new view and in that sense a literature review is a reinterpretation of the research problem. Figure 5 illustrates one model using the idea of literatures and the notetaking categories. All filters apply to all papers.

Coverage This paper belongs to the ___ literature.

Synthesis

• What does this paper say?

• What has not been said, or done, or discussed?

• How is this paper situated in the history of the field and this topic?

Methodology What are main methods and research techniques? Advantages? Disadvantages? What claims are made? Are these research methods appropriate to support the claims?

Significance What is the practical significance of the research? What is the theoretical significance of the research?

Figure 4. Possible categories for notes for a literature review

	Literature Area			
	P1	P2	P3	etc
Note Contents filtered through a category	What does this paper say?	What does this paper not say?	Main methods used in the paper	Significance of this paper
Literature Review	A literature review in four parts: What does this paper say, not say, what methods are used, and what is the significance of this literature area?			

Figure 5. A model of literatures and note taking categories used to create a literature review

An explanation of the model in Figure Five

From a literature, published papers (P1, P2, P3, etc) are identified, located, read, and noted. Note categories are decided on. In Figure 5, the note categories are what this paper says, doesn't say, the main methods used, and the significance of the paper. These categories can act as filters. For each paper, take notes using these categories. Using the same notetaking categories for all the papers makes it easier to create a synthesis.

After you have used these categories to note several papers, you can begin to write your literature review. Using your answers to the first question (*What does this paper say?)* you have notes from all your papers answering this question. Pull those answers together and you have a synthesis of answers to the first question. Continue with the other categories. These notetaking categories come from Boote and Beile (2005); if these categories are not appropriate, create categories that are. The categories should make sense in terms of your own purpose and research questions.

What if nothing has been written on my topic?

With thousands of articles published every year, it is doubtful if absolutely nothing has been researched, written, and published on your topic. However, assume that nothing has, or at least you cannot find anything. It is not enough, however, to simply state that no publications exist for at least two reasons. First, it is not likely that anyone will believe that nothing has been written on a topic, given that so much has been published on so many subjects. *Gatekeepers*, a way of referring to those in charge of accepting or rejecting research papers, including journal editors, journal reviewers, conference reviewers, and course instructors, will probably challenge such a claim. Second, it is impossible to prove a negative; you cannot prove that something does not exist.

If you wish to make the claim that no publications exist on your topic, you will have to demonstrate clearly that you have searched the literature and come up empty-handed. To do this, your first step would be, as plainly as possible, to define your topic so readers all know what you were looking for. Next, you should list the databases you searched and the descriptors you used. Then, tell us what you found and why those findings do not fit your topic. Then, identify related fields. For example, if you are working in applied linguistics, related fields might be psychology, sociology, literacy, developmental education, and communication studies. Explain why these fields might contain publications that could inform your topic. Then, examine each of these fields in the same why you examined your primary topic area. Demonstrate that you have carefully searched these areas by telling us briefly but exactly what you searched, how you searched, what you found, and why these finding do not apply to your topic. One of two things will probably happen: you will have convinced the gatekeepers and other readers that you have diligently searched several literatures and found nothing, or, and this is to be preferred, you will have found many relevant publications in the process. Either way, you have performed a literature search and have much to write about.

The relationship of Literature Review, Purpose, and Research Questions

It would appear when reading a typical journal article that RQs come from the literature review,

since if one reads a paper from the beginning, one reads the literature review before reading the research questions. However, probably the reverse is true. The literature review is inspired by and must support the research question, not the other way around.

Where can we say so far?

A complete introduction to a research paper includes five parts: a general topic indictor, a statement of the specific aspect of the topic or a statement of the problem, a literature review, the purpose of the paper, and one or more research questions (see Swales, 1990; 2004). It is not always necessary to follow this order strictly. For example, if the literature review is long, it might be good to put a preliminary purpose before the literature to remind readers of your research purpose while they are reading the long literature review. You do not want readers asking themselves halfway through your literature review, "Why am I reading this?" Another approach is to spread the literature review throughout the entire paper (Boote & Beile, 2005), in which case a research question might be the final product of a paper using an ethnographic design rather than a starting point of a paper using an experimental design.

Purpose and Research Question(s)

A statement of purpose and one or more research questions (RQs) typically follow the literature review. Since the purpose statement is often only a paragraph and a single research question is often only one sentence, it is easy to ignore the significance of this critical section. There is no rule as to how many RQs to have, but most papers have two or three. On the other hand, a thesis or dissertation may have five or six since it is typically a book-length work.

To grasp the importance of the purpose and research question(s), imagine a playground in which there is a teeter-totter (also known as a seesaw), which is a long board balanced over a fulcrum. A child sits on one end and balances another child on the other end, and together they push themselves up and down. The purpose and RQ function as the fulcrum balancing the question (statement of the problem, literature review) and the answer (method, results, discussion). The power of this image is that it illustrates that the purpose and RQ not only balance the question and answer, but they connect them. For example, a literature review is justified by the purpose and RQ in that it is the purpose and RQ that help the researcher decide what in the literature review to include and exclude.

One example of a complete introduction

To illustrate, here is an example of an introduction from Griffee (Unpublished) showing in bold the title, general topic or "why this topic is interesting to the field," statement of the problem, literature review, purpose, and the research question.

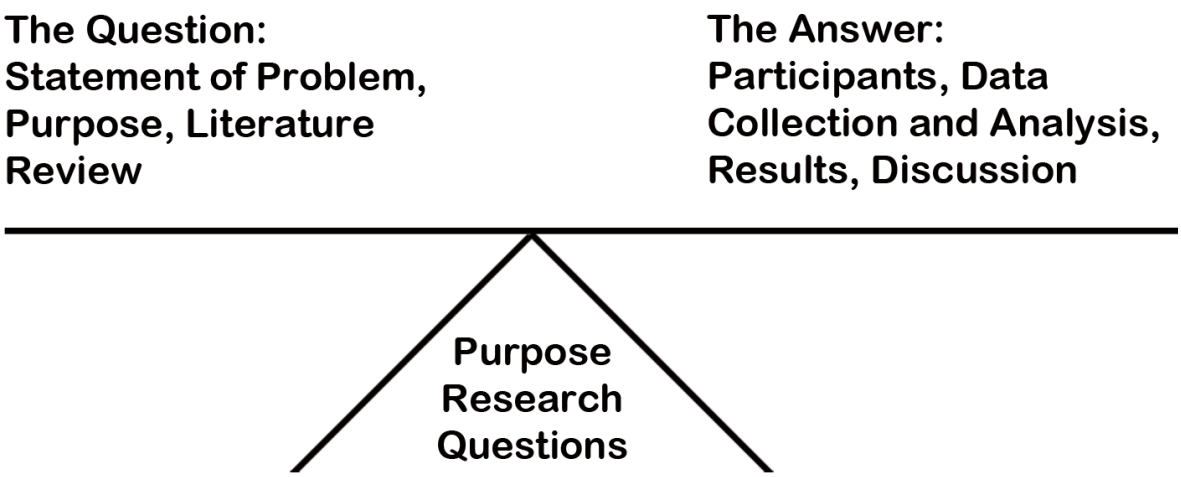


Figure 6. The relationship of literature review, purpose and research questions, and body of the research paper

Title Student Estimation of High School Preparation in Researching and Writing an Academic Research Paper

General Topic and/or why interesting to the field. Because of its basic position in the curriculum, Freshman English has been described as one of the most demanding, challenging, and complex of educational experiences (Feinstein, 1950). **Statement of the problem.** One of the major challenges of first year English is an assignment to write an academic research paper (ARP). Nevertheless, many college instructors believe their students have not been adequately prepared in high school to write a research paper. **Literature Review.** For example, Fletcher (1963) claims that high school libraries are not adequate to support a full term paper. Taylor (1965, p. 127) maintains that first-year students are not equipped to carry on meaningful literary research, and that high school English teachers are not prepared to offer expert guidance and instruction. His image of the ARP is that of students "regurgitating the thoughts of others" and of "stringing together of quotations." Relying on interview data, Schwegler & Shamoon (1982) support the claim that ARP writing is a factual information gathering exercise. Gaston, Smith, and Kennedy (1995) say that entering freshmen have fuzzy knowledge of research, distortions, and misimpressions from high school. They observe that students do not see the point of research because they have never been exposed to research.

Baird (1992) claims "overwhelming and widespread apathy in today's high schools" (p. 66) relative to writing in general and the academic research paper specifically. In an informal survey of 100 of his freshman students, Carino (1984) estimated that about 60% of his students had written high school term papers, which he characterizes as comprising: 10 pages in length, a narrative summary of an overly broad topic, not limited in scope, often a biography on the life of great person, and

containing non-acceptable sources, for example, magazine and encyclopedia articles. He concludes, "[T]he research-writing experience that students gain in high school does not prepare entering students to write the research paper required in freshman writing courses" (Carino, 1984, p. 6). **Purpose.** The purpose of this article is to discuss the nature of preparation high school students perceive they received researching and writing an academic research paper in high school. Developmental educators would be better able to structure classroom class activities if they not only knew the experiences that shape student's present perceptions towards writing, but also knew how these perceptions can be responsible for current writing problems. **Research Question** The specific research question is, do students in a first year college English course who are researching and writing an academic research paper rate their high school training as adequate preparation for that task?

Participants

The *Participants* section of a paper describes the persons involved in the study. Typically they are students in your class, often the whole class, but depending on the purpose, research questions, and practical considerations, participants could also be a single student, a small group of students, all the students who visited the writing center this semester, or even very large groups of persons whom you do not know and have never met, such as everyone in this year's incoming freshman class.

The purpose of the *Participants* section is to describe the individuals in the study in enough detail that the reader can understand who they are. The question the reader has is, "Are these participants similar enough to my students so that the results of this study might apply to them?" If the participants are not described, readers cannot answer this question, thus the conclusions cannot be generalized to other situations. The research design could be mentioned and discussed in the *Participants* section to the extent it affects how participants are selected.

Some of the questions for you to consider in describing your participants are: the number of participants, their ages, genders, when and where they were investigated, their school levels, a course description, the day/hour of course meetings, their majors, language proficiency, why the participants were selected, their academic background, whether they are graduate or undergraduate students, their language background, their socioeconomic statuses, their places of origin, their class attendance, data on who dropped out, who the researcher was, whether the researcher was also the teacher, (and if so, was the researcher the only teacher involved)? If there were other teachers involved, who were they, how were they selected and involved, and what did they do? Reporting on your role as researcher is a fairly new development, but is now done because in a post-positivistic world the researcher is not seen as an outside, uninvolved observer. Rather, the researcher can and does play a role, so researchers describe their own roles in the process.

How many questions about participants must you answer? There is no specific number, but you want to answer enough questions for readers to "see" your participants. The criterion for any category of information is, "Does this descriptor fit my research question?" For example, if your research question involves the family life of your students, then you need to describe participant's

family. If your purpose and research questions do not involve family life, then probably such information would be superfluous.

Finally, you should consider how you will gather participant data. All teachers know some of the basics about their students, such as gender and attendance records, but may not have other information such as whether students live with their families or in dormitories. You do not want to find yourself writing your research paper after the class is finished only to discover certain information about your students is needed to discuss and explain your results, but that you do not have this information. By that time, it might be difficult or even impossible to locate students to ask for it. To avoid this situation, decide early on what information you need, and devise a procedure to gather it. This might include having students write something about themselves and asking their permission to use this information in a study; conducting interviews; or writing a questionnaire for students to complete.

Materials

The purpose of the Materials section is to describe materials used in the study and explain their function. For most teachers, materials are usually data collection instruments such as questionnaires, tests, observation schemes, compositions, diaries, homework, or interviews, but materials could also include equipment such as audio recorders or video cameras. Equipment should be described; data collection instruments, space permitting, should be included in the appendix so readers can see for themselves what participants used.

Data collection instruments should be described in detail, including how they were prepared and piloted; how reliability was calculated; and how they were validated. If an instrument was a test or questionnaire, what was the scoring method? At this point, don't explain how materials were used--this belongs under Procedures. Similarly, don't give actual scores--these belong under Results.

To summarize:

1. List each material and piece of equipment in the order it was used, or list materials by research question.
2. Tell how it was prepared.
3. What was its purpose? For example, which research question did it answer?
4. How was it piloted?
5. Give reliability estimates derived from the pilot. Later, in the Results section, provide the reliability for the actual results in your study. Each time an instrument is administered, reliability should be calculated since it is not a universal, but a characteristic of a set of particular scores (Thompson, 2003).
6. Provide scoring or rating scales.
7. Tell us how the instrument was validated.

Procedures

For each piece of equipment or each data collection instrument, explain how it was used and what the participants did. For example, if a questionnaire was administered, readers want to know how many students actually filled it out, what were the environmental conditions (room, time of day), and how long it took to complete the questionnaire. It might be a wise idea to begin and maintain a log in which this type of information could be written down and dated. The issue in the method section is replication. In other words, does a reader have enough information from your report so that, if they wanted, they could do the same thing you did?

Analysis

The analysis section has various names including data analysis, design, and statistical procedures. The purpose of the analysis section is to tell readers how the data were analyzed in order to answer the research questions. In your analysis section, it would be appropriate to state the design you used, which analysis you employed to answer each research question, why you used that particular analysis, and what it will tell you.

State the Research Design. If you haven't already, state what research design you are using and give a citation so your readers can read more about it. Common research designs are experimental designs and quasi-experimental designs, but case study and survey design are also well known. The research design was probably first mentioned in the literature review, and possibly mentioned again in the Participants section. If the research design impinges on the analysis, state how that impingement works. For example, in an experimental or quasi-experimental design, probably the control and experimental groups will be compared and the resulting data analyzed, and how that data will be analyzed can be explained here.

State which Data Analysis you used. To report your data analysis, consider your research question or questions, and ask yourself how you plan to answer each question. By "answer each question," I mean what kind of data collection instrument you plan to use, what kind of data you plan to get from it, how you plan to analyze the data, and how the results of that data analysis will answer your research question. If a statistical procedure is performed, for example, to compare pre-test to post-test scorers of experimental and control classes, then report the statistic you plan to use, why that particular statistic was selected, what its assumptions are, and what you hoped to accomplish by using it. In the following examples, two scenarios are described in regular text; the actual analysis is in **bold** text.

<u>Example one</u>

Your colleague Amanda teaches two classes of developmental English. Last semester, in one class she taught the way she usually does--she lectured on the genre organization and writing plan. In the other class, she tried a new brain-based technique she read about in Smilkstein (2003), in which she workshopped both the genre organization and writing plans. Amanda decided to use a quasi-experimental design with a control and experimental class because she wanted to compare her two classes. Her research plan was to grade the final composition in both classes, and if the treatment class scored higher than the control class, she would argue that her innovation

worked. Her data analysis will be a statistical test called a *t*-test, which is designed to measure statistical differences between sets of scores. Her research question is, "Will students in a basic writing course using a brain-based curriculum increase their writing ability as measured by final test scores compared to students in a course using a traditional curriculum?" Her analysis section might read as follows:

A quasi-experimental post-test-only design was employed using a control and experimental group. Final compositions from both classes were scored and assumptions for a t-test were checked including equality of variance. An alpha p-value of .05 was set.

Example two

An ESL teacher, Griffee (1993), wondered how the conversations his students studied in textbooks compared with natural conversations between native speakers (NSs). He found various ways to analyze conversations. Then he found two textbook conversations and constructed roleplays on the same topics. He asked two native speaker colleagues to read the roleplay cards and have a conversation on the topic while he placed a tape recorder on the table. The research question is posed as a purpose: Despite the many arguments advanced in favor of the use of authentic materials, not many empirical studies have emerged to show what using authentic materials contributes to classroom learning. The aim of this paper is to consider that question by examining to what extent the conversational dialogs in textbooks correspond to unplanned conversations in structured roleplays.

Two NS role plays were recorded, transcribed and compared with two textbook conversation dialogues according to ethnographic components of communication (Saville-Troike, 1989, p. 138) and linguistic components of discourse analysis (LoCastro, 1987; McCarthy, 1991; Cook, 1989). See the Appendix for a full transcription of all four conversations.

Results

In the Results section, you report what you found, being careful only to report, avoiding the temptation to discuss your results. It is helpful to present your results in the same order as your research questions. In some cases, tables may help to summarize the data. Brown (1988, p. 54) recommends a prose description followed by a table, but in chapter four, Swales and Feak (2004) seem to suggest a table followed by a data commentary. Whichever organization you choose, you should use it consistently in your paper.

If you are reporting scores from tests or results of a statistical analysis, report N-size, *p*-value, test reliability for group scores being reported and the Standard Error of Measurement (SEM) if individual participant scores are relevant, magnitude of effect or effect size, and descriptive statistics of the scores in the study being reported. For a discussion of null hypothesis statistical significance testing including N-size and *p*-values, see Griffee (2004). Finally, it is helpful to report confidence intervals of the statistical procedures used.

For results of qualitative data gathering, do not include raw, unanalyzed data unless you wish your readers to have access to it. In that case, include it in an appendix. The reader already knows

the data collection instrument you used to gather the data (materials), how you went about it (procedures), and how you analyzed it. In the Results section, we want to know what you found.

Discussion

In your introduction, you posed one or more research questions; now it is time to answer them, preferably beginning each answer with a "yes" or a "no." If the Discussion section is short, many TREEs elect to combine it with the Results section or substitute the word "Conclusion" for the discussion. Either way, this is where you tell your readers what it all means. We know from the Results section *what* happened, now we want to know *why* it happened. After you interpret your answers for us, think about their implications. Perhaps you could suggest changes in practice, reflect on unanswered questions, or suggest directions for future research.

References

A reference is a short citation consisting of the basic descriptors that typically identify books, journal article, magazines, newspapers, reports, brochures, or even movies and songs. In short, a reference is a locator device for anything that might be mentioned or discussed in your paper. References are placed at the end of a paper in an agreed-upon format; the two most common (in the humanities and social sciences) are APA (American Psychological Association) and MLA (Modern Language Association). Both the APA and MLA are large academic organizations, which in addition to holding major conferences, sponsor journals that publish articles of interest to their members. Over time, these organizations created format rules that eventually grew into publication style manuals. Many academic disciplines including applied linguistics, developmental education, and English as a Second Language (ESL) use APA formatting.

The reference section is directly connected to each cited source throughout the paper, but especially to the literature review. If we see the literature review section as telling someone about your friends, the reference section is what provides your friends' addresses and contact information. The reference section is important because it holds all the citations for the paper, and as such functions as the paper's database. In that sense, the reference section is part of the paper's external validity or generalizability, because it is part of the replication process. Without the reference section, a researcher who wanted to replicate all or part of your research would not be able to do so because he or she would not have access to the same background information. In the reference section, the author provides, free of charge so to speak, a mini database, a treasure house of sources. For this reason, when reading research articles, the reference section of those articles should be read and carefully searched for additional material that can contribute to your research project.

Using the correct form tends to be a technical but persistent writing problem for many TREEs (and experienced writers!). The correct form depends on the style manual you use. For example, in the APA style, a single author article published in a journal gives the author's last name followed by his or her initials, the year of publication, the title of the article, the name of the journal which is in italics, the volume number also in italics, the issue number in parentheses, and the page numbers. Like everything else in research, nothing is value-free, and even style manuals may involve an implicit epistemology (Madigan, Johnson, & Linton, 1995). Notice in this example that

the second line is indented, usually five spaces.

Griffee, D. T. (2002). Portfolio assessment: Increasing reliability and validity. *The*

Learning Assistance Review, 7(2), 5-17.

A major editing task in writing a paper is to match internal paper citations with those in the reference section. Go through you paper and for each citation, check it off in the reference section. Even though publication style manuals discuss all aspects of writing a research paper, some issues are more important than others in the sense that they are more common. See the appendix for some of the citation styles you may find yourself using frequently.

Appendix

Citation/Reference Guide based on APA (6th Edition)

Submitting a paper Justify the left hand margin only. Remember to check that all citations in the paper are listed in your references, and all citations in the reference section are in the paper. Submit your papers double-spaced because single spacing does not allow enough room for comments.

In the body of the text

One author citation

In discussing the relationship of evidence, Brown (1988) states . . .

In a study of evidence (Brown, 1988), it was found that . . .

Citation with several authors

Several studies (Gomez & Jones, 1979; Smith, 1988; Griffee, 1999) have noted . . .

Ary, Jacobs, and Razavieh (1990) compared . . .

Author & page number (giving page numbers is recommended for books & long articles)

It is strongly suggested (Dunkel, 1990, p. 70) that . . .

It is strongly suggested (Dunkel & Gorsuch, 1990, p. 70) that . . .

Dunkel and Gorsuch (1990, p. 70) strongly suggest that . . .

Reference section at the end of the paper

Journal articles

Griffee, D. T. (2002). Portfolio assessment: Increasing reliability and validity. *The Learning Assistance Review, 7*(2), 5-17.

Books

Casazza, M. E., & Silverman, S. L. (1996). *Learning assistance and developmental education*. San Francisco, CA: Jossey-Bass.

Edited Book

Jones, S., & Smith, C. R. (Eds.). (1986). *Bilingual education*. New York, NY: Praeger.

Article or chapter in an edited book

Griffee, D. T. (1997). Where are we now? Trends, teachers, and classroom research. In D. T. Griffee & D. Nunan (Eds.). *Classroom teachers and classroom research* (pp. 23-35). Tokyo: Japan Association for Language Teaching.

For a chapter in a book that is not edited, include "In" before the book title.

Table formatting

Table 1

Title of Table Underlined with Key Words in Caps but No Period

content of the table between the lines here

Figure formatting

Insert the figure.

Figure 1 in (italics). In the caption, capitalize only the first word and proper names. If the caption takes more than one line, double-space and continue aligned left.

What about Qualitative Research Paper Structure?

The term *qualitative research* is an imprecise term that creates at least two problems. One problem arises because *qualitative* refers to a type of data, not a type of research design. This makes the idea of qualitative research illogical--even impossible since it ignores the possibility of triangulation using multiple types of data, some of which might be qualitative and some of which might be quantitative. Furthermore, the term assumes the type of data determines the type of research. Nevertheless, despite these problems, many researchers and editors continue to use the term *qualitative* to refer to a type of research instead of restricting its use to a type of data. A second and related problem is that the general term *qualitative research* can refer to any number of research designs, for example case study, ethnography, or grounded theory. In these research designs, none of the assumptions of experimental design, such as variables and hypothesis testing are operative. Can the research paper structure described in this chapter be used with so-called qualitative research? Will the structure of the classic research paper accommodate research that does not use variables and answer pre-specified research questions, but instead seeks to follow investigative themes and see what emerges?

At this point, there are at least two solutions to this dilemma. One approach is to use the research paper structure described here. The other approach is to find or create a structure

more compatible with your research design. Shohamy (2004) suggests the second answer. She questions the use of tightly defined categories such as *method* and *discussion* because she wants to keep research open and flexible to allow for innovation. On the other hand, it may be possible for you to use the paper structure described in this chapter, referred to as the standard research paper structure, because this structure is elastic and accommodating. For example, the Teachers of English to Speakers of other language (TESOL) guidelines for reporting critical ethnographic studies states: "Because of the diversity of perspectives represented within ethnography, be as explicit as possible about the disciplinary traditions or models of ethnographic scholarship that have influenced your work" (Chapelle & Duff, 2003, p. 173). This comment seems to refer to a literature review. If there were any way to use the classic research paper structure, the advantage would be that it has been around for a long time and many of the organizational and procedural problems have been discussed and worked out.

If, however, you decide you cannot use the standard research paper form, your option is either to find a structure similar to your research design by reading in the literature, or to use the TESOL guidelines to create one yourself (Chapelle & Duff, 2003). The advantage to finding or creating your own structure is that your reporting form fits your research; the disadvantage is that it makes your job twice as hard. We are gradually moving toward not only multiple forms of data, but perhaps multiple forms of research paper writing, but we are not there yet. The TESOL guidelines mentioned earlier show that this transition is in the process of being worked out, but the same guidelines also indicate that no consensus has yet been reached.

DISCUSSION QUESTIONS

Write down one or two questions you had while reading this chapter about research paper structure.

Task 1. What was the easiest section in this chapter for you to understand?

Task 2. What was the most difficult section in this chapter for you to understand?

Task 3. What do you think makes it difficult?

References for Structure of a Research Paper

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Baird, S. E. (1992). Writing with meaning in a dispassionate age. *English Journal, 81*(8), 66-68.

Boote, D. N., & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educational Researcher, 34*(6), 3-15.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Carino, P. A. (1984). The college research paper vs. the high school research paper. *Indiana English, 7*(2), 6-11.

Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly, 37*(1), 157-178.

Cook, G. (1989). *Discourse*. Oxford: Oxford University Press.

Feinstein, G. W. (1950). What are the aims of freshman English? *College English, 11*(8), 459-461.

Fletcher, P F. (1963). Should term papers be abolished? *The Clearing House, 38* (September), 32.

Gaston, T. E., Smith, B. H., & Kennedy, R. P. (1995). Putting the research paper into rhetorical context. In J. Ford (Ed.). *Teaching the research paper: From theory to practice, from research to writing* (pp. 55-67). Metuchen, NJ: The Scarecrow Press.

Griffee, D. T. (1993). Textbook and authentic dialogues—What's the difference? *The Language Teacher, 27*(10), 25–33.

Griffee, D. T. (2004) Research in practice: Understanding significance testing program evaluation. *Journal of Developmental Education, 27*(3), 28-34.

Feinstein, G. W. (1950). What are the aims of freshman English? *College English, 11*(8), 459-461.

LoCastro, V. (1987). Aizuchi: A Japanese conversational routine. In L. Smith (Ed.). *Discourse Across Cultures: Strategies in World Englishes* (pp. 101-113). New York, NY: Prentice Hall.

Madigan, R., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist, 50*(6). 428-436.

McCarthy, M. (1991). *Discourse analysis for language teachers*. Cambridge: Cambridge University Press.

Saville-Troike, M. (1989). *The ethnography of communication*. (2nd ed.) London: Basil Blackwell.

Schwegler, R. A., & Shamoon, L. K. (1982). The aims and processes of the research paper. *College English, 44*(8), 817-825.

Shohamy, E. (2004). Reflections on research guidelines, categories, and responsibility. *TESOL Quarterly, 38*(4), 728-731.

Smilkstein, R. (2003). *We're born to learn: Using the brain's natural learning process to create today's curriculum*. Thousand Oaks, CA: Corwin Press.

Swaffar, J., & Bacon, S. (1993). Reading and listening comprehension: Perspectives on research and implications for practice. In A. O. Hawley (Ed.). *Research in Language Learning: Principles, processes, and prospects* (pp. 124-155). Chicago, IL: National Textbook Company.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.

Swales, J. M., & Feak, C. G. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed). Ann Arbor, MI: The University of Michigan Press.

Taylor, T. E. (1965). Let's get rid of research papers. *English Journal, 54*(2), 126-127.

Thompson, B. (2003). Guidelines for authors reporting score reliability estimates. In B. Thompson (Ed.). *Score reliability: Contemporary thinning on reliability issues* (pp. 91-101). Thousand Oaks, CA: Sage.

Tuckman, B. W. (1999). *Conducting educational research* (5th ed.). Orlando, FL: Harcourt Brace.

PART TWO
Design

INTRODUCTION TO RESEARCH DESIGNS:

WHAT EXACTLY IS A RESEARCH DESIGN?

It is often hard to think in the abstract about how to conduct research. Teacher-Researcher-Educator-Evaluators (TREEs) often want clear categories in order to be able to think about research; if you can't think clearly, it makes it more difficult to act clearly. It helps to familiarize yourself with research categories because it gives a sense of "what to do." Two terms helpful in thinking about research are *design* and *data*, or more specifically, *research designs* and *data collection instruments*. In Part Two of this text, titled Design, we will discuss various research designs. In Part Three, titled Data, we will discuss various data collection instruments.

The question and the answer

An introduction to a research article, starting from a statement of the general topic and concluding with a specific research question or questions, can be seen as stating the question the author intends to ask. In that sense, the elements of the introduction are necessary for any research article, in that it would be hard (but maybe not impossible) to imagine a research paper that did not clearly state its purpose somewhere close to its beginning. How we answer the question, on the other hand, is the function of the design. This can vary greatly in form depending on the design chosen.

THE QUESTION	THE ANSWER
Introduction	Research Design
(Purpose and RQs)	

Figure 1. The structure of a research paper

What is research design?

Creswell (2003, p. 13) uses the term *strategy* to explain design. He also uses the terms *tradition, method, approach, procedure*, and *process*. Ary, Jacobs, and Razavieh (1990) say a research design is "a description of the procedures to be followed in testing the hypothesis" (p. 110). For Babbie (2004), research design "involves a set of decisions regarding what topic is to be studied among what populations with what research methods for what purpose" (p. 112) Gay and Airasian (2000) define research design as "a general strategy for conducting a research study" (p. 107), which describes the basic structure of the study. They explain that research design tells the researcher which strategy to select, and includes the hypotheses, variables, and real world

constraints. Vogt (1999) in his dictionary of statistics and design defines research design as "[t]he science (and art) of planning procedures for conducting studies" (p. 247). In other words, a research design is a set of instructions for data collection and analysis. As examples of research design, Vogt (1999) gives experiments and quasi-experiments. Finally, for Mohr (1996, p. 104), a research design is the basis for causal conclusions.

A research design can be defined as an operating model or blueprint for a research project, which accounts for internal reasoning (causality) and external reasoning (generalizability). If the introduction in a research paper raises the question to be investigated, the research design contains directions to determine the answer. As a blueprint is to building a house, a research design is to conducting a research project. The research design stipulates the parts of the research project, how they are arranged, and how they function. However, the research design does not determine the type of data, how the data are collected, or how they are analyzed any more than a blueprint tells a house builder the color of the walls or what furniture will occupy its various rooms.

How many research designs are there?

To answer this question, I examined the tables of contents from as many research methods books as I could find. I did not consider specialized texts such as those discussing only one type of approach or those discussing one design exclusively. Invariably, such introductory research textbooks contain a section of what Creswell (2003) and others variously call design, strategy, tradition, method, approach, procedure, process, and orientation. I interpret these terms to be or point in the direction of research designs. A convenience sample of twenty textbooks was examined: Ary, Jacobs, and Razavieh (1990); Babbie (2004), Bernard (1994); Brown and Rogers (2002); Charles and Mertler (2002); Cohen, Manion, and Morrison (2000); Creswell (2002; 2003; 2005); Gay and Airasian (2000); Heppner, Kivlighan, and Wampold (1999); Johnson (1992); Lodico, Spaulding, and Voegtle (2006), McDonough and McDonough (1997); McKay (2006); Nunan (1992); Neuman (2000); Paltridge and Phakiti (2010); Seliger and Shohamy (1989); and Tuckman, (1995).

It is illuminating to list research design chapters by their frequency of mention, that is, how many of the methods books listed here contain a chapter on a certain design. Such a listing of chapters on various designs may point to a level of consensus on the acceptance of the design, and perhaps to the level of perceived importance attributed to various research designs.

Of the 23 possible research designs listed in Table 1, can we be sure all of them are actual research designs? For example, #4 is correlational design. In the research methodology textbooks I examined, I found eight examples of chapters listed as correlation. This suggests to me that correlation is a way of conducting research, and thus a research method. But correlation is a statistical procedure--an example of data analysis. Therefore, the list of 23 designs may actually be shorter. Nevertheless, we can say that if a TREE were to pick up a research methods textbook, there is a high chance of finding a chapter on experimental and survey design, a medium chance of finding a chapter on case study or action research design, and a relatively low chance of finding a chapter on narrative or grounded theory design.

Table 1. Research Designs from Tables of Contents listed by frequency of mention

Research Designs	Frequency of Mention	Percent of Mention
1. Experimental	19	.95
2. Survey	16	.80
3. Ethnographic	9	.45
4. Correlational	8	.44
5. Case Study	8	.40
6. Action Research	8	.40
7. Qualitative	6	.33
8. Ex post facto	4	.22
9. Descriptive	4	.22
10. Introspection	3	.17
11. Grounded Theory	3	.17
12. Narrative	3	.17
13. Historical	3	.17
14. Evaluation	3	.17
15. Causal-Comparative	1	.06
16. Interactional analysis	1	.06
17. Phenomenology	1	.06

Other designs that were mentioned but not given full chapter status include:

18. Critical Theory, Heppner, Kivlighan, and Wampold (1999, p. 240)

19. Constructivism, Heppner, Kivlighan, and Wampold (1999, p. 238)

20. Content analysis, Neuman (2000, p. 150)

21. Ethnomethodology, Babbie (2004, p. 290)

22. Feminist Research Design, Neuman (2000, p. 82); Heppner, Kivlighan, and Wampold (1999, p. 240)

23. Hermeneutics, Heppner, Kivlighan, and Wampold (1999, p. 239)

How to select a research design

One way to decide on a research design is to select one that addresses your purpose and answers your research questions. This seems to be common sense. If you are clear about your research questions, you can ask yourself which design could best answer them, and select that design. The problem with this approach is the hidden assumption that either there is only one answer, or that there is a best answer. In fact, any design can answer your research questions. Asking which design could answer your research question would be like asking which car in a group of cars could take you from your home to your school. The answer is that, unless they are defective in some way (or out of fuel), they all could. Nevertheless, asking which design best answers your purpose could work.

Second, you can use the design you already know. Perhaps most of the research papers that you have read use a certain design, and all or most of your professors, colleagues, or other research mentors use this design. In a strange way, this makes sense, even if it is limiting. We tend to use what we know, and if we only know one design, then that is the one we will select.

A third and related approach is to use the design promoted by a senior colleague. Many researchers believe that only one design, which incidentally happens to be the one they prefer, is legitimate. These researchers often influence their colleagues—especially those who work under them--to use the same design. Using a design that a senior researcher suggests can be a political decision. It also has the advantage that the professor under or with whom you are working can give cogent advice on the research because of familiarity with the design.

Fourth, it might be that your academic specialty has a preference toward doing a research using a certain design. Since you want to gain admission to this area of specialization, it makes sense to use this design. A variation of this approach is that you want to publish in a certain journal and it seems to publish a certain type of research using a certain type of design. Therefore, you might use this design to increase your chances of publication. Again, this is a political and pragmatic decision rather than an academic one, but one that is understandable.

Fifth, you might be intrigued by a certain design and want to learn how it works. Therefore, you choose this design to guide your research. By doing so, you broaden your area of expertise. This is a research-design-as-adventure point of view, which has many benefits.

These are some of the reasons, both political and practical, why researchers pick a research design; there may be more. It is hard to say that some of these reasons are better than others. The only rationale that I would advise against is picking a design thinking it to be easy. There is no such thing as an easy design. Nonetheless, even if you pick a design for that reason, you can still learn a lot--including the fact that it will probably not be as easy as you first thought.

Mixed-method design

One might hear a lot about mixed-methods design (Green, 2007; Smith, 2006), but it is not always clear what the term means. One problem lies in the ambiguity of the term *method*, which means a way of doing something. However, what exactly one is doing is not specified or included in the

term itself. *Method* could refer to a way of organizing research (in which case the term *design* is preferable), it could refer to a way of collecting data (*data collection instrument* is better in this case), or it could refer to a way of analyzing data. Often the term *method* is used without definition. In the preface of their book on mixed methods, Tashakkori and Teddlie (1998) use the terms *method* or *methodology* 27 times without defining it. Nor do they include either term in their index. *Method* is frequently used in close proximity to other terms such as *paradigm*, *quantitative*, and *qualitative*, but these terms are seldom if ever defined (see for example Angouri, 2010). If by *method* one means a method of research or research design, then a mixed-method design creates a logical perplexity. Can there be an experimental survey or an ethnographic experiment? This is difficult to imagine since each design has its own logic of organization, which cannot be mixed or combined with other designs.

A careful reading of the literature suggests that the term *mixed methods* means the use of more than one kind of data (Barkaoui, 2010). If this is the case, the term *mixed-method design* may not be appropriate because it simply means more than one type of data is being collected. Given the current popularity of triangulation or multiple data collection, most research could be called mixed-method design. Such a wide application of the term runs the risk of becoming meaningless, or at least not very informative.

Let's see how mixed methods might work. Tashakkori and Teddlie (1998) cite a model of mixed methods based on Patton (1990, p. 188) in which Patton considers *design* (subdivided into naturalistic and experimental), *measurement* (subdivided into quantitative and qualitative), and *analysis* (subdivided into content and statistical. By combining the subcategories, Patton arrives at six categories, and argues for what he calls multiple forms of triangulation. The results can be seen in in Table 2.

Table 2. Examples of mixed methods as a result of triangulation based on Patton (1990)

Design	Measurement	Analysis
1. Experimental	Quantitative	Statistical
2. Naturalistic	Qualitative	Content
3. Experimental	Qualitative	Content
4. Experimental	Qualitative	Statistical
5. Naturalistic	Qualitative	Statistical
6. Naturalistic	Quantitative	Statistical

As listed in Table 2, design type one would be experimental, using quantitative

data, and statistical analysis of those data. This is the classical experimental design, or what Patton calls the pure deductive approach. Design #2, naturalistic inquiry, uses qualitative data and content analysis. This is pure (or at least typical) inductive approach, which Patton terms qualitative strategy. But types #3, #4, #5, and #6 could be called mixed forms. For example, design #3, experimental, uses qualitative data and content analysis. There is no reason the mixing could not be expanded. For example, experimental design using both quantitative and qualitative data with appropriate analysis, or naturalistic design (a case study design or a grounded theory design) again with quantitative (for example, scores from a standardized test) and qualitative (maybe interview transcripts) data and analysis appropriate to the data.

We need to look carefully, however. If the term *method* is defined as data collection and data analysis, we have examples of mixed data *methods*, but not mixed *designs*. Nevertheless, Patton's model helpfully illustrates the possibility of using multiple forms of data and analysis within a design. Whether or not it is necessary to give these combinations a new name, such as mixed methods or mixed model studies, is open for discussion.

Is there such a thing as qualitative research?

As in the case of mixed-method design, one hears the terms *qualitative* and *quantitative* research, but what do these terms actually mean? Is there such a thing as a qualitative research design, or quantitative research design? I would argue *no*. Consider this example: Suppose a teacher investigating his/her own class elects a case study research design. In addition, scores from a norm-referenced, standardized test of some sort were available. Such scores are potentially valuable because they might indicate whether this group of students is relatively the same (homogeneous) or relatively different (heterogeneous) on whatever construct the test purported to measure. If we stopped here and only used quantitative test data, this case study might be termed quantitative research. But suppose we also want to include observation and interview data, which is typical for case studies. Is there any problem with combining qualitative observation and interview data with quantitative test data in the same case study design? Not at all. Now we can see the problem: Any design can accommodate any type of data. To identify research by the type of data gathered is not only confusing and unhelpful, it is by and large not possible if multiple types of data are gathered, which is frequently done.

What if this same teacher used only qualitative data? In that case could we say it is qualitative research? We could, but doing so would not tell us much because the term *qualitative research* conceals more than it reveals. Instead of using the term qualitative or quantitative research, let us come down a level of abstraction and identify the research design we propose to use.

What conclusions can be drawn about research designs?

Keeping in mind that the research method textbooks examined here may not be a representative sample, some tentative conclusions can still be drawn. First, there appear to be many research designs available to researchers, but there is no agreement as to the exact number. Second, various research designs are inconsistently named. These inconsistencies seem to be examples of what Pedhazur and Schmelkin (1991) call the "Jingle Jangle fallacy." The Jingle fallacy states that because *constructs* are called the same name, they are the same; the Jangle fallacy states that

because constructs are called by different names, they are different. In other words, just because authors call things by the same name does not mean they are the same, and just because authors call things by different names does not mean they are different. An example of using the same name but with a different meaning is Ary, Jacobs, and Razavieh (1990) and Gay and Airasian (2000), who both use the term *descriptive.* On close inspection Ary, Jacobs, and Razavieh (1990) mean descriptive statistics while Gay and Airasian (2000) mean survey. An example of using different names with the same meaning is Ary, Jacobs, and Razavieh (1990) and Gay and Airasian (2000) who both use the term "causal-comparative" while Cohen, Manion, and Morrison (2000) use the term "*ex post facto*." Both terms seem to be pointing to the same design.

Finally, as shown in Table 1, some research designs are mentioned more frequently than others. In the research methods books examined, experimental design was given chapter-level discussion in all but one book. On the other hand, several research designs are given chapter-level status by only one or two authors. This suggests that perhaps these designs are not popular or not favorably viewed, at least by the North American research methodologists who wrote the majority of the books. We might conclude that if a TREE were to elect to use a popular research design, such as a variation of experimental design or survey design, for a paper, thesis, or dissertation, the decision would be met with understanding and probably approval (Lazaraton, 2005). On the other hand, if a TREE elected a less-popular research design, such as action research or grounded theory, the decision might be met with less understanding. This is not to say that lesser-known research designs should not be selected, but it does suggest that gatekeepers, such as professors and journal editors, may not be as familiar with these designs. As a result, the TREE may have to explain the selection of these designs more fully. The lesson is to carefully consider design selection, remembering that the decision may be as much political as it is academic.

References for Introduction to Research Designs

Angouri, J. (2010). Quantitative, qualitative or both? Combining methods in linguistic research. In L. Litosseliti (Ed.), *Research methods in linguistics*. New York, NY: Continuum.

Ary, D., Jacobs, L., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Orlando, FL: Harcourt Brace.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thompson.

Barkaoui, K. (2010). Do ESL raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.

Bernard, H. R. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). Walnut Creek, CA: Altamira Press.

Brown, J. D. & Rogers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.

Charles, C. M., & Mertler, C. A. (2002). *Introduction to educational research* (4th ed.). Boston, MA: Allyn and Bacon.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Merrill Prentice Hall.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (2nd ed.). Upper Saddle River, NJ: Pearson.

Gay, L. R., & Airasian, P. (2000). *Educational Research: Competencies for analysis and application* (6th ed.). Upper Saddle River, NJ: Merrill, Prentice Hall.

Green, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Heppner, P. P., Kivlighan, D. M., Jr., & Wampold, B. E. (1999). *Research design in counseling* (2nd ed.). Belmont, CA: Wadsworth.

Johnson, D. M. (1992). *Approaches to research in second language learning*. New York, NY: Longman.

Lazaraton, A. (2005). Quantitative research methods. In E. Hinkel (Ed.). *Handbook of research in second language teaching and learning* (pp. 209-224). Mahwah, NJ: Lawrence Erlbaum.

Lodico, M. G., Spaulding, D. T., Voegtle, K. H. (2006). Methods in educational research: From theory

to practice. San Francisco, CA: Jossey-Bass.

McDonough, J., & McDonough, S. (1997). *Research methods for English language teachers*. London: Arnold.

McKay, S. L. (2006). *Researching second language classrooms*. Mahwah, NJ: Lawrence Erlbaum.

Mohr, L. B. (1996). *The causes of human behavior: Implications for theory and method in the social sciences*. Ann Arbor, MI: The University of Michigan Press.

Neuman, W. L. (2000). *Social research methods* (4th ed.). Boston, MA: Allyn and Bacon.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Paltridge, B., Phakiti, A. (Eds.). (2010). *Continuum companion to research methods in Applied Linguistics*. London: Continuum International Publishing.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Pedhazur, E. S., & Schmelkin, L. P. (1991). *Measurement design and analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Seliger, H. W., & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.

Singer, M. (2005). *The legacy of positivism*. New York, NY: Palgrave Macmillan.

Smith, M. L. (2006). Multiple methodology in education research. In J. L. Green, G. Camilli, P. B. Elmore, A. Skukauskaite, & E. Grace (Eds.), *Handbook of complementary methods in education research* (pp. 457-475). Mahwah, NJ: Lawrence Erlbaum.

Tashakkori, A., & Teddlie C. (1998). *Mixed methodology: Combining qualitative and qualitative approaches*. Thousand Oaks, CA: Sage.

Tuckman, B. W. (1999). *Conducting educational research* (5th ed.). Orlando, FL: Harcourt Brace.

Vogt, W. P. (1999). *Dictionary of statistics & methodology* (2nd ed.). Thousand Oaks, CA: Sage.

CHAPTER THREE

SURVEY RESEARCH DESIGN (SRD)

Surveys have broad appeal, particularly in democratic cultures, because they are perceived as a reflection of the attitudes, preferences, and opinions of the very people from whom the society's policy makers derive their mandate. (Rea & Parker, 1992)

In this chapter you will learn the basic concepts of sampling that are necessary to understand survey research. This chapter will not discuss data collection instruments commonly used in Survey Research Design, such as questionnaires, observations, or interviews, because these instruments can be used with other designs. You will also be introduced to many key terms associated with SRD listed in the glossary at the end of this chapter.

Questions to think about

1. Have you ever taken a poll or a survey?

2. Do you generally trust the results from polls and surveys?

3. Have you heard about a recent poll or survey? Explain its focus and what it was trying to determine.

Introduction

Survey research is well known and widespread in many countries. Hardly a day goes by without reading the results of a survey in the newspaper, being asked to take a survey on the Internet, or hearing about the results of the latest poll on T.V. In the social sciences, survey research has experienced a dramatic increase (Berends, 2006). Smith and Davis (2003) confirm that, in experimental psychology at least, surveys and questionnaires have not only survived, but have thrived and are popular in current research. For a short history of survey research, see Neuman (2000, p. 247). In the second language field, a major problem continues to be the over-identification of questionnaire instruments with survey design. Even experienced researchers such as Rea and Parker (1992) identify survey research as an example of data collection--which it is not. A survey is a research design, while a questionnaire is a data collection instrument. Since any research design can accommodate data from any collection instrument, a questionnaire and the data it produces can be used in any research design and should not be exclusively identified with survey design research.

Survey Research Design (SRD) defined

A survey design uses various data collection procedures to enable the teacher-researcher-educator-educator (TREE) to investigate a construct by asking questions of either fact (descriptive)

or opinion (explanatory) from a sample of a population for the purpose of generalizing to the population. The term *survey* is an umbrella term that allows for many data collection procedures including questionnaires, interviews, and observations.

What are its historical roots and beliefs?

Surveys have been used since ancient times for census and tax purposes. Charles Darwin and his cousin Francis Galton used questionnaires in the late 1800s to gather support for their theories on evolution and intelligence (Goodwin, 2003). Surveys, as we have come to know them, date from the 1930s (Ary, Jacobs, & Razavieh, 1990, p. 410). George Gallup founded his American Institute of Public Opinion in 1935 (Rea & Parker, 1992).

Survey design studies purport to measure a construct which may be theoretical (for example, opinion, beliefs, attitudes) or practical (for example, ownership of certain objects, time spent on certain tasks). Survey design studies, especially those providing opinions, gather data made by respondents after the fact. As a result, the data is in one sense subjective and unverified and in another sense objective. For example, if we ask a certain group how many hours a week they spend talking with their families and they report a specific number, subjectively we may wonder if that number is correct, but objectively, it is the number they reported, and we can work with it (Bernard, 1994, p. 261).

Key components of Survey Design

Ary, Jacobs, & Razavieh (1990, p. 411) list five steps in the survey process: Planning the survey, sampling or deciding whom to survey, constructing the instrument, conducting the survey, and analyzing the data. Brown (2001, p. 8) lists six similar steps: planning the survey, developing the instrument, gathering the data, analyzing the data statistically, analyzing the data qualitatively, and reporting the results.

How survey design might look as a visual

A survey design consists of a construct, a population of interest, a sample, and a data collection instrument used to measure the construct in the sample.

The construct must be identified and defined, the population must be described, and the sampling process must be explained. A *construct* is the term TREEs give to what they are researching. A construct may be well defined due to previous research, or it may be ill defined and vague. But well defined or vague, the construct must be stated; otherwise, how could a researcher sample a population on an unstated issue, problem, or concern? The construct-defining process often takes place in the literature review and is reflected in the purpose section. The construct plays an important role when it comes to selecting and constructing the data collection instruments.

A *population* may be persons or things, but more often than not in education and applied linguistics it comprises persons. In addition to defining the construct, survey research must explain why we would reasonably expect a population to possess this construct.

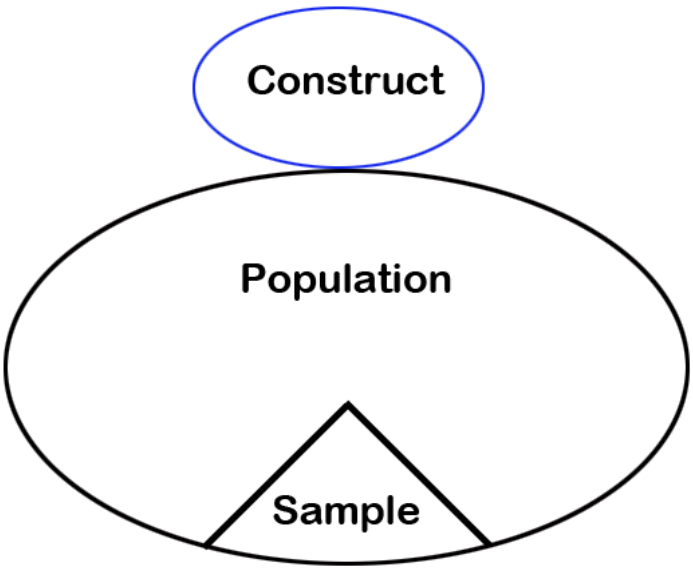


Figure 1. Three important components of any survey design: construct, population, and sample

A *sample* is part of the population that the TREE will survey using some type of data collection instrument, often a questionnaire or observation protocol. This process of collecting data from a smaller group and generalizing interpretations about this group to a population is called *sampling*. The sample must be shown to be representative of the population. Survey design data collection instruments such as questionnaires and observation tend to ask carefully-constructed-but-answerable questions rather than profound-but-hard-to-answer questions; profound questions concentrating on a few persons would be better accommodated by another instrument, perhaps interviews.

A brief scenario

Greta teaches educational linguistics in a university second language studies program. For some time, she has been interested in the extent to which classroom teachers use lectures to present material. Recently, a policy recommendation was issued by the state educational commission which stated that when possible, student-centered classroom procedures should be used, and one of the specific directives was the use of pair and group work to replace lectures as a way of introducing and working on material. The teachers Greta talked to in her school and at state conferences indicated that they agreed with this new policy, but Greta wonders what other teachers think and to what extent they are actually implementing the recommendation to use lecture format less often and engage in tasks more. Because this is both a state and local issue, she wants data from teachers statewide, school-wide, and department-wide. She decides to embark on a study using a survey design with a questionnaire to investigate teacher opinions and

practices across the state. In addition to her survey, she will employ an observation technique to investigate school practices, and finally interviews with teachers in her department.

What is Survey Design good at?

Survey design can produce not only descriptive summary, but also generalized statements based on large databases (Cohen, Manion, & Morrison, 2000, p. 171). Survey design is good for providing information for curriculum development, including needs analysis and program evaluation, as well as researching certain topics (Brown, 1997). It is also good at collecting data on groups too large to observe directly (Babbie, 2004). Data collection instruments, especially data from questionnaires, can be used to gather data in a relatively short period of time. This data can be analyzed statistically and result in generalizable results. For this reason, a complex survey can be used as the basis for a thesis or dissertation. In fact, surveys and their resulting data and interpretations can serve as the basis for a book. For example, Jarvis (1991) interviewed and surveyed university academics for a book on achieving tenure. He was able to expand sections of his questionnaire to sections in a book.

What is SRD not good at?

Survey design does not attempt to establish cause-and-effect relationships. In addition, even though surveys are good at collecting descriptive data, or learners' feelings and opinions, they are not good at directly measuring learning that has or has not taken place. Nor is survey design good at providing detailed explanation; surveys tend to be a mile wide and an inch deep. For this reason, surveys have trouble dealing with complexity and subtlety; people's opinions don't always fit the data collection instrument used in a survey. Finally, survey instruments, especially questionnaires, typically can't change in midstream to reflect new insight.

What kind of data is most closely associated with SRD?

The type of data collected in a survey is directly related to and controlled by the purpose and research questions of the particular study. Rea and Parker (1992) discuss three types of data typically used in SRD: description, behavior, and preference. *Descriptive data*, also known as *demographic* data, refers to information such as respondent's age, gender, place of birth, level of education, and ethnicity. *Behavioral data* refers to actions that respondents do, such as hours spent talking to family, number of books read in a year, transportation choices to work or school, and the like. *Preference*, also known as *opinion*, refers to perspectives respondents hold, such as what they think about certain aspects of classroom instruction. SRD data gathering instruments can include face-to-face interviews with individuals or groups, telephone interviews, or written questionnaires (Brown, 1997) as well as direct observation, and composition ratings (Johnson, 1992). Depending on the data collection instrument, surveys allow for the collection, analysis, and interpretation of either quantitative or qualitative data.

What issues and concerns does survey design highlight?

There are four issues that any TREE considering the Survey Research Design must ask:

1. *What am I investigating?* The issue is identifying the construct.

2. Who will be surveyed? The issue is identifying the various levels of population including the working population, and the sample frame.

3. How should I select my respondents? The issue is selecting the type of sampling for your survey.

4. How many respondents should I survey? This issue is generally known as sample size.

What am I investigating? As in any type of research design, survey research begins with identifying and defining the construct, the purpose, and the research questions. All of these are crucial, if for no other reason than these three elements taken together strongly influence which data collection instruments to use. To deal with this, you have to identify the research construct, and how it influences the survey purpose and research questions. Much research begins in some way or other with *theory*, and SRD is no exception. Theory is a term that covers many things including "interrelated concepts/constructs, definitions, and propositions to provide a systematic view of social phenomena" (Berends, 2006, p. 626). Even if a TREE has no explicit theory when first beginning a survey, he or she has reasons for considering the survey; he or she wants to investigate something. The TREE needs to think as carefully as possible about this, because the operationalization gets translated into purpose, then into research questions, and then finally to instrument items.

Who will be surveyed? It may be that the persons you wish to survey are already known to you; in fact, it may be that as soon as the idea for a survey popped into your mind, a certain group of persons was immediately present. Two caveats come to mind: first, for some TREEs it may not be obvious who or what to survey; second, exactly how to identify this group may not be obvious. For these two reasons, it may be helpful to think about what is called *population*. Survey researchers (Rea & Parker, 1992; Chromy, 2006) have developed some systematic ways of thinking about and obtaining a survey population.

One of the first, if not the first step, is to consider the *target population*. This is a vaguely defined, almost theoretical group. Usually we do not know who exactly is included in this group, or where they are. Examples of target populations are "all ESL students," "undergraduate students studying Japanese," or "international teaching assistants (ITAs)".

A next step entails deciding on the survey *unit of analysis*. A unit of analysis could be students, libraries, textbooks, teaching assistants, teachers, schools, departments, graduate status (undergraduates, graduates), males, females, ethnic groups, or any combination (for example, Chinese teaching assistants). Any of these could be surveyed; in choosing which one, you are picking a unit of analysis and also narrowing your target population.

The next step is to identify a *general population*. General population is a vague concept, and TREEs engaged in survey research seldom know who constitutes this population. Suppose the unit of analysis is international teaching assistants. ITAs arrive and graduate on a regular basis. How many schools use ITAs, and in which departments? To answer these questions, we have to define

a *working population*. A working population is the operationalization of the general population. Here we try to make or gather a list with individual names. Here is also where *bias* can become a problem. Bias is the systematic exclusion of a group from our general population as we create a working population. Suppose for example, we ask a department in our university for a list of the persons for our research, and we get a list of all students in our working population who live in dorms. This list, however, excludes all those persons who do not live in dorms because, for example, they are married. In other words, we have created a bias against married students that systematically excludes them from our survey. We wish to create a list of persons (assuming people were our unit of analysis) that is called a *sampling frame,* defined as the list from which we will draw our sample.

How should I select my respondents? To deal with the third issue in respondent selection, you have to consider which type of sampling to use. The type of sampling used in a survey determines the level of generalization that can be claimed by the researcher. We assume you have constructed a sampling frame that is a list of all known, and therefore possible, respondents. The question is, given this list, how do you select which persons on the list to include in the sample?

Another way of putting the question is: *How do you like my pizza?* Suppose I make a pizza and ask you to come to my home for a taste. I give you a slice and ask, "How do you like my pizza?" What *you* don't know is that half my pizza has meat toppings and half has vegetarian toppings. What *I* don't know is that you have a strong preference for one of the toppings, but not the other. Without looking, I offer you a random slice of my pizza. You register a strong response, good if I picked the topping you like and bad if I did not. Based on your response, I conclude that you like (or don't like) my entire pizza when in fact the slice you received was not representative of the whole pizza. Had I given you another slice from another part of the pizza, your response might have been different. Because the slice of pizza I gave you was not representative of the whole pizza, for you to make an informed decision about my pizza, I need to make sure that the slice of pizza that I offer you is representative of the whole pizza. This is an example of *sampling*. In discussing types of sampling, I will discuss *probability sampling* and *nonprobability sampling*, with four examples under each type as shown in Figure 2.

In the pizza metaphor, the pizza is a working population and the slice I gave you is a sample. In a valid survey, any slice I might give you should be representative of the whole pizza so that your response to the slice would represent your response to the whole pizza. Staying with the pizza example, the first issue, *what are the construct, purpose, and RQ?* could be stated as, *Should we have a pizza?* The second issue, *what is the population of interest?* could be stated as, *What kind of pizza should we have?* The third issue, how to select respondents, can be framed as, *Is this a typical slice?* The fourth issue, how to determine sample size is reframed as, *how big does the slice need to be in order for you to decide if you like the whole pizza?*

Nonprobability sampling is defined by Rea and Parker (1992) as sampling in which the TREE does not know the probability of a given possible respondent's being selected into the sample. It is generally agreed that nonprobability sampling is widespread (Kalton, 1983; Warner, 2008). Four types of nonprobability sampling will be mentioned: *convenience, purposeful, snowball*, and *quota* sampling.

Type of Sampling	Characterized as
Nonprobability sampling	Subjects selected by the researcher
1. Convenience	A group already formed and easy to use
2. Purposeful	Knowledgeable and available persons
3. Snowball	Selected respondents suggest other respondents
4. Quota	Stratified sampling, but not randomly chosen
Probability sampling	Subjects selected by a random mechanism
1. Simple random	Pull names out of a hat
2. Systematic random	Computer generated numbers to select
3. Stratified	The sample divided into groups called strata
4. Cluster	Groups of strata

Figure 2. Types of sampling with working definitions

Convenience sampling is, as the name implies, a sample that is easy or convenient for the TREE to find. Examples include intact classes, participants at a conference workshop, or volunteers. In other words, the group is there and willing to participate.

Purposeful sampling is a technique where the TREE decides who would most likely be of help in informing us about our construct. In the case of international teaching assistants, that could be department heads that employ many ITAs, academic advisors who supervise them, ESL instructors who teach them, and undergraduate students who study under them.

Snowball sampling is when the TREE can identify a few qualified respondents, survey them, and then asks them if they can identify additional respondents. Each one-find-one, or better yet, each one-find-several, would be the motto of a snowball sampler. For a description of how researchers used snowball sampling to find a highly specialized group, heroin addicts who recovered without treatment, see Biernacki (1986).

Finally, *quota* sampling is when the TREE decides that the sample needs to include a certain quota of persons because, taken together, they represent the population. For example, imagine a certain population--all ESL students in your state. The TREE believes that the male/female ratio is 60-40 and the ethnicity is half Chinese, a quarter Koreans, and a quarter “other”. As a result, the TREE makes sure 60% of the sample is male and 40% is female. Furthermore, the TREE makes sure Chinese students make up half the sample, Korean students a quarter of the sample, and “others” account for the remaining quarter. But none of the participants is chosen randomly.

The TREE, in looking for Chinese students, surveys all the Chinese students he can find on a first-come-first-served basis until he achieves the desired number.

Probability sampling is defined by Rea and Parker (1992) as "the probability of any member of the working population being selected to be a part of the eventual sample is known" (p. 147). Four types on probability sampling are mentioned: *simple random, systematic, stratified*, and *cluster* sampling.

Simple random sampling refers to the process by which all possible names (sampling units) are identified (sampling frame) and selected in a random, arbitrary way. If the number of possible respondents is small, the classic way of random selection was to write each name on a small piece of paper, put them all into a container of some sort, and have a disinterested person who could not see into the container pull out the desired number. For a larger number, a list could be made and assigned, and then selected in a random way, such as every nth number. The advantage of simple random sampling is that it is relatively easy to do and is effective if the population is homogeneous (relatively the same); the disadvantage is that the population may be heterogeneous (relatively different being composed of many subgroups) in which case the TREE is not sure to what extent all groups are represented.

Systematic random sampling is used when the sampling frame is large, for example, over 100,000, which makes pulling names out of a hat impossible. *Systematic* refers to using a computer program to generate numbers and select the desired number required in a sample. This type of sampling is useful for TREEs engaged in dissertation-level survey research.

Stratified random sampling is the division of a population into exclusive and exhaustive units called *strata* (Levy & Lemeshow, 1991). Stratified random sampling is used when the working population is heterogeneous; that is, the working population contains groups of interest that are of unequal size. Because of their unequal sizes, smaller groups might not be accounted for in simple or systematic random sampling. Typical examples of such groups (strata) are gender, ethnicity, L1, and age. Suppose a TREE is interested in Chinese and Bengali speaking ESL students. The TREE estimates that the working population includes about 50% Chinese speakers and 10% Bengali speakers. Simple random sampling would probably include enough of the Chinese speakers for the sample, but may not select enough Bengali speakers. To ensure the sample includes the ratio of Chinese and Bengali speakers the TREE wants to include, she identifies Chinese speakers and Bengali speakers as strata. In that way, separate samples are selected from each group (Kalton, 1983). Members from each stratum are randomly sampled in about the same proportion as identified in the population (50% and 10%). "Operationally, a stratified random sample is taken in the same way as a simple random sample, but the sampling is done separately and independently within each stratum" (Levy & Lemeshow, 1991, p. 99). In this way, you can be reasonably sure that both groups will be represented in your sample in about the same proportion as you estimate they are represented in the population.

Cluster sampling is typically used for surveys involving very large populations, over 100,000, and is a version of simple random sampling, with the units of analysis being groups or clusters instead of individuals. Cluster sampling is used for very large populations because lists of

individual units of analysis may not be found; in fact, they may not even exist--for example, all 3rd grade ESL students in the state of Texas. This list may not exist because K-12 schools are grouped in school districts. The first cluster may be identified as all school districts in the state. From this list, a sample may be made. From that sample, all elementary schools may be identified--at least the number if not the school names. From this list, the number of 3rd grade classes may be identified, and from this list a sample may be drawn. Cluster sampling is a way of narrowing down a large, vaguely defined population to a more manageable group so a list can be developed from which a sample may be drawn.

Sample size: How many respondents should I survey? We now turn to the fourth issue, sample size. For the remainder of this discussion, only two terms will be used: *population* and *sample*. Population refers to the working population from which sampling frame or list of possible respondents was made. Reference to population will be to the group of persons of interest in the survey; it is the group the TREE wishes to generalize to. Sample, on the other hand, will mean the group of actual respondents selected from the population by one of the sampling procedures discussed. Returning to the pizza metaphor, the question is, *How big does the slice have to be in order for you to decide if you like the pizza or not?* In other words, do you have to eat all the pizza before deciding, would half the pizza be enough, or would less than half be enough, and if so, how much less than half? Sample size is a major concern: if a sample is too large, it becomes unfeasible to engage the survey; if the sample is too small, then the results cannot be generalized to the population. *Generalized to the population* means the results of the survey sample can be applied to the population, and generally means having a sample size large enough to generate a 95% or 99% confidence interval in terms of scores of plus or minus 3, 5, or 10 (Rea & Parker). Therefore, a sample size that generates that level of confidence is important.

Brown (1988) reminds us that, at least when it comes to sample sizes, large is good because the larger the sample size, the more it resembles the population. Unfortunately, there is no way to say exactly what large means. Brown (2001) offers three helpful pieces of advice. First, he reminds us that in statistics, when the N-size is in the neighborhood of 25 to 30, the distribution is at or approaching normal, and normal distribution is a requirement for the statistics used to determine generalizability. Therefore, it is unlikely that any sample size lower than 25 to 30 would be adequate; unfortunately, a sample size of 30 does not always guarantee an adequate sample size. Second, if the language program involved in the survey is small, one solution would be to survey everyone. The good news would be that the sample size issue has disappeared; the bad news would be that any findings could be generalized to the language program and probably nowhere else. Brown's (2001) third piece of advice is to suggest a survey sample be at least 50. Depending on the number of participants in the language program, a sample size of 50 is more likely to be adequate than one of 30 because it is more likely to result in a sample size that would be normally distributed, but even with these numbers, we cannot be sure.

Combining the insights found in Brown (2001) and Rea and Parker (1992), we can make reasonable estimates of what an adequate sample size would be. Figure 3 shows three common survey situations and possible ways of estimating a sample size.

Population situation	Population estimation	Sample population
My class	5 - 60	The entire class
A language program	61 – 100	50% should be enough.
A research project	101 - 10,000	A sample size of 400 can likely produce generalizable results.

Figure 3. Three sample size estimates

The first situation, termed *My Class,* represents a single intact class with an enrollment between five and sixty members. This is a convenience sample that is available to the TREE, but the results cannot be generalized because we have no way of knowing if or to what extent this group of students resembles the larger population of interest. As Warner (2008, p. 75) states, "At best, when researchers work with convenience samples, they can make inferences about hypothetical populations that have characteristics similar to those of the sample." I interpret that to mean that a convenience sample might function as a pilot study to assist in defining the target population. A convenience sample might also help validate a data collection instrument, such as a questionnaire. Nunan (1992) illustrates this possibility:

> Small-scale studies may decide to use non-probability samples because they are easier to establish and, in consequence, cheaper. They may also be perfectly satisfactory for a preliminary or pilot study whose aim is to trial survey instruments and procedures, not to obtain data which can be generalized from sample to population. (p. 142)

It might also be possible to calculate the mean and standard deviation of this convenience sample for the purpose of estimating the mean and standard deviation (called standard error of the mean) of the population. Finally, the TREE could use a convenience sample to make a weak claim.

The second example is a language program with a population of 61 to 100, possibly of combined classes, or possibly each student in the program. Perhaps the TREE is interested in research or perhaps in program evaluation. If the latter, then a generalization to the language program is not a problem because it is the language program that is the population of interest. Thus, for a population of 100, a sample size of 50 seems appropriate. Even if normal distribution cannot be obtained, Rea and Parker (1992, p. 133) state that for "population sizes for which the assumption of normality does not apply, the appropriate sample size is 50% of the population size."

The third situation is an estimated population size of 100 to 100,000. By using the chart from Rea and Parker (1992, p. 133), it is possible to estimate the minimum sample sizes for small population up to 100,000 (but see Fowler, 1993). The TREE estimates a population, and then selects a confidence interval (plus or minus 3%, 5%, or 10%) all at 95% level of confidence. As an example, for an estimated population of 100,000, the minimum sample size is 1,058 for 3%, but only 383 for 5%, and a mere 96 for 10% at the 95% level of confidence and sample sizes of 1,809, 659, and 166 at the 99% level of confidence. Of course, surveys suffer from low response rates, so a more prudent sample size would be more than the minimum. For an estimated population of 5,000, the minimum sample size is 880 for plus or minus 3%, 357 for plus or minus 5%, and 95 for plus or minus 10% all at the 95% level of confidence and 1347, 586, and 161 for the 99% level.

Conclusion

We can make some general statements about SRD.

1. SRD does not establish a cause and effect relationship. In general, establishing cause and effect relationship is precisely the job of research designs, but SRD is one exception.

2. SRD attempts to establish a relationship between a sample and the population from which the sample is drawn.

3. SRD uses a statistical procedure to establish this relationship.

4. No other design uses statistical procedures in this way; in fact, no design uses statistical procedures to establish cause and effect. Statistical procedures alone do not establish cause and effect (although they may be used to provide evidence for such a relationship).

Further Reading

A survey reprinted from *The Journal of Educational Research* can be found in Gay and Airasian (2000, p. 304-314). See Gorsuch (2000) for an example of a nationwide survey of Japanese teachers using sophisticated analysis techniques. Brown (2001) included a survey he administered to members of TESOL, and Meskill, Anthony, Hilliker-Vanstrander, Tseng, and You (2006) report a statewide survey on CALL uses and preferences. Most introductory research textbooks contain a chapter on survey design—for example, Ary, Jacobs, & Razavieh (1990), Bernard (1994), Cohen, Manion, & Morrison (2000), Gay & Airasian (2000), and Johnson (1992). In addition, some books are devoted entirely to survey research data collection procedures: Bourque & Clark (1992), Brown (2001), Converse & Presser (1986), de Leeuw, Hox, & Dillman (2008), Fowler (1993), Fox & Tracy (1986), Holstein & Gubrium (1995), Kalton (1983), and Rea & Parker (1992), In Applied Linguistics, Brown (2001) or Dörnyei and Taguchi (2009) would be required reading for any serious work with surveys.

DISCUSSION QUESTIONS

Note some questions you had reading about survey design.

Find a partner, Ask for their questions and tell them yours. You can read your questions as part of the class discussion. Alternatively, you can write your questions on the board so all can share the insights and use them as a whole class discussion.

Reflection on Survey Research Design

1. What is the attraction of Survey Research Design for you?

_____

_____

2. What problems or issues would you anticipate in using SRD?

_____

_____

_____

Task 1. Find a published research paper that either implicitly suggests or explicitly states that the researcher used a survey research design. Answer the following questions and report to the class.

1. State the title of the article, the author, the year, and the name of the journal in which the paper was published.

2. What construct does the survey measure?

3. Are the purpose and the research question (or questions) clearly stated?

4. What in the paper tells you the design? In other words, do you agree that SRD was used?

5. Was a population defined?

6. Was the sampling procedure made clear?

7. What is your overall opinion of this article? Could it serve as a model for a survey you might like to do?

Task 2. If you decide to conduct a survey, how would you deal with these issues?

The construct:

_____

_____

The population:

_____

_____

Sample selection:

_____

_____

The sample size:

_____

_____

The data collection instrument:

_____

_____

Glossary of Key Survey Research Design terms

Bias Warner (2008) defines bias as, "A systematic difference between the characteristics of a sample and a population" (p. 4). For example, if a certain population, such as all of the students in your language school, are composed of 70% men and 30% women, but your sample has 50% men and 50% women, your sample is biased against men and biased towards women; men are underrepresented in your sample and women are overrepresented. In research design, bias is not a personal feeling towards certain groups, but a structural problem in the study's design.

Census A survey, often done by a government, to establish baseline data for any question of interest. Example areas of interest for a national government might be number of persons in a household, educational level of each person in the household, level of income, and how many persons are married, divorced, or single. What separates a national census from a large survey is that the census attempts to include all its citizens whereas the population of a survey tends to be limited and more narrowly defined.

Cohort A cohort is a group of people that have certain characteristics of interest to the researcher. For example, a cohort might be those learners born in a non-English speaking country but who immigrated to an English-speaking country before the age of 17. A cohort study would identify such a group and survey them at, say, high school graduation, then identify members of the same population (but not the same individuals) five years later, and perhaps survey them again in another five years.

Confidence Interval A measure of "the level of sampling accuracy that the researcher obtains" (Rea & Parker, 1992, p. 126). In Survey Research Design, Confidence Intervals are also known as the *margin of error*. Ayres (2007) offers a shortcut to determine the confidence interval he calls the *two standard deviation* rule (2SD). His shortcut is easy because it only requires the mean and standard deviation. Imagine a test with a scale of 0 to 100. You obtain a mean of 70, and a standard deviation of 8. You can be 95% confident that the true score (the score of the population) is two standard deviations plus or minus the mean. Using the 2SD rule, we can be 95% confident that our population mean is between 54 and 86.

Confidence level The amount of uncertainty in a study a TREE is willing to accept. Most researchers choose a 95% level of confidence resulting in a five percent chance of error.

Construct A mental or psychological representation of a quality or trait that we wish to measure. Every DCI (Data Collection Instrument) presupposes a construct. If someone were to ask you what you were measuring with, for example, a questionnaire, whatever answer you gave, that answer would be the construct. Constructs are not physical; the brain is an organ, but the mind is a construct. Examples of typical constructs of interest in Educational Linguistics are motivation and proficiency.

Descriptive A term that is sometimes used in place of the term survey research (Gay & Airasian, 2000). Descriptive data, also called *demographic data*, is data describing respondents by such variables as age, gender, and ethnicity.

Explanatory The ability of a research design to explain what we want to know. For example, if a TREE is interested in why certain groups of respondents do or think something, questionnaire data in a survey design might be helpful in explaining why because of the ability of questionnaire items to capture respondent opinion.

Generalizability The ability to transfer or apply conclusions reached by studying a sample population to a larger population. Brown (1988, p. 113) defines generalizability as "the degree to which the results can be said to be meaningful beyond the study."

Interview A face-to-face encounter between a researcher asking questions and a respondent answering the questions. Interviewers follow a line of questioning called a *protocol*, and may vary from strict (all questions known in advance, asked exactly as stated, and asked the same order to all respondents) to rather loose (a general question is asked with the answers and follow up questions arising from the first question).

N-Size N stands for number. The N-size is the number of participants in a group, a sample or the general population.

Nonprobability sampling Any type of sampling that allows the TREE to select some persons for inclusion in a sample without allowing others. In other words, the nonprobablistic sample does not guard against bias.

Observation Another word for looking. In research, observation is intentional and systematic rather than casual and random. Although the classroom is frequently the object of observation, any situation can serve as a data-gathering scene for an observation.

Panel A panel is defined as a longitudinal survey of the same people over time (Creswell, 2002). Suppose we are interested in the opinion of a certain group of students on certain topics, and we conduct our survey the year they graduate, then a year later, and then five years later. This group of students would be a panel.

Population The group the TREE is interested in. In everyday language, a population refers to all the persons living in a defined place, but for research, population has a more restricted meaning. A population is a group that is defined in some way; typically a group of persons (but it could be groups, schools or even objects) with certain characteristics of interests, for example, all the students in a certain university taking second language courses or all U.S. students studying German. The population is sometimes referred to as the general population to distinguish it from the *working population* and the *sample population*.

Population parameter Certain characteristics of the population of interest that could be used in defining the population or in creating the sample. For example, if the population of interest is all the students in North America who are studying French, a population parameter might be undergraduate (or graduate), traditional student (or returning student), male (or female), etc.

Probability sampling Any type of sampling that allows any member of the sample population to be selected for the sample, as opposed to nonprobability sampling, which does not.

Questionnaire A data-collection instrument that asks respondents for demographic information, opinion or questions of fact. Questionnaires typically ask respondents to quantify their answer by circling a number (say, one to five) thereby providing numerical data that can be statistically analyzed. Alternatively, questionnaire items may be open ended and provide qualitative data.

Respondents Persons who are solicited to answer interview questions or fill out questionnaires; people who respond to our inquiry.

Sampling The process of selecting and surveying a small portion of a larger group. The assumption is that the sample has the same type of persons (assuming people are the *unit of analysis*) as occur in the larger population. To the extent that this is not the case, the survey suffers from *sampling error*.

Sampling error The difference between the mean and standard deviation of the population and the mean and standard deviation of the sample.

Sampling frame A list from the working population of possible members of the general population. The sampling frame is used to eventually identify the actual sample.

Sample population or sample size The group that will be actually studied; the group to whom the TREEs will survey. This sample population has a mean and a standard deviation. The standard deviation of the sample population is also termed the *standard error*. Generally, the larger the sample size, the smaller the error.

Sample statistics "The statistics calculated from the survey sample can be used to make inferences about the larger populations be they teachers, schools, or districts" (Berends, 2006, p. 625).

Strata Plural is strata; singular is stratum. Associated with a type of sampling called random, stratified sampling. Strata are identified subpopulations of interest such as males, females, students who have studied abroad and students who have not. After identifying the population, Brown (2001, p. 73) suggests identifying "salient characteristics of the population" commonly known as *strata*.

Trend A change over time in any population or subpopulation. For example, are persons getting married at a later age, the same age as ten years ago, or are persons getting married at a younger age. The answer can be determined by a survey at a certain time. If the answer continues in the same direction at a later time, say X years later, this can be identified as a trend.

Unit of analysis Used to describe and analyze the survey population. The population is made up of aspects of interest to the researcher, for example, teachers, students, university departments, schools, households, courses, computers, graduating classes, and the like. These aspects are called units and are referred to as the unit of analysis. Unit of analysis is the same as the *unit of assignment* used in Experimental Research Design terminology.

Working population Defining the working population is an intermediate step between describing the general population and the sample population. The population of a study, for

example, all ESL students studying in North America, might be a population of interest; however, this population is largely unknown to the researcher because ESL students come and go. To narrow this general population down (operationalize) into a more manageable group, a working population is defined. The idea is to be able to generate as complete as possible a list of possible members of the general population. "This list is known as the *sampling frame*, and it is from this list that the actual sample is eventually drawn" (Rea & Parker, 1992, p. 141).

References for Survey Research Design

Ary, D., Jacobs, L., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Orlando, FL: Harcourt Brace.

Ayres, I. (2007). *Super crunchers*. New York, NY: Bantam Dell.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thompson.

Berends, M. (2006). Survey methods in educational research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.). *Handbook of complementary methods in education research* (pp. 623-640). Mahwah, NJ: Lawrence Erlbaum.

Bernard, H. R. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). Walnut Creek, CA: Altamira Press.

Biernacki, P. (1986). *Pathways from heroin addiction recovery without treatment*. Philadelphia, PA: Temple University Press.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (1997). Designing surveys for language programs. In D. T. Griffee & D. Nunan (Eds.). *Classroom teachers and classroom research* (pp. 109-121). Tokyo: Japan Association for Language Teaching.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge University Press.

Bourque, L. B., & Clark, V. A. (1992). *Processing data: The survey example*. Thousand Oaks, CA: Sage.

Chromy, J. (2006). Survey sampling. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.). *Handbook of complementary methods in education research* (pp. 641-654). Mahwah, NJ: Lawrence Erlbaum.

Converse, J. M., & Presser, S. (1986). *Survey Questions: Handcrafting the standardized questionnaire*. Thousand Oaks, CA: Sage.

Cohen, L,, Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

de Leeuw, E., Hox, J. & Dillman, D. (2008). *International handbook of survey methodology*. The European association of methodology book series. Florence, KY: Taylor & Francis, Psychology Press.

Dörnyei, Z. & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. London, UK: Routledge.

Firebaugh, G. (1997). *Analyzing repeated surveys*. Thousand Oaks, CA: Sage.

Fowler, F. J. Jr. (1993). *Survey research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. Thousand Oaks, CA: Sage.

Gay, L. R., & Airasian, P. (2000). *Educational Research: Competencies for analysis and application* (6th ed.). Upper Saddle River, NJ: Merrill, Prentice Hall.

Goodwin, J. C. (2003). Psychology's experimental foundations. In S. F. Davis (Ed.). *Handbook of research methods in experimental psychology* (pp. 3-23). Malden, MA: Blackwell.

Gorsuch, G. (2000). EFL educational policies and educational cultures: Influences on teacher's approval of communicative activities. *TESOL Quarterly, 34*(4), 675-710.

Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Thousand Oaks, CA: Sage.

Jarvis, D. K. (1991). *Junior faculty development; A handbook*. The Modern Language Association of America. New York: Author.

Johnson, D. M. (1992). *Approaches to research in second language learning*. New York, NY: Longman.

Kalton, G. (1983). *Introduction to survey sampling*. Thousand Oaks, CA: Sage.

Levy, P. S., & Lemeshow, S. (1991). *Sampling of populations: Methods and applications*. New York: John Wiley.

Meskill, C., Anthony, N., Hilliker-Vanstrander, S., Tseng, C., & You, J. (2006). CALL: a survey of K-12 ESOL teacher uses and preferences. *TESOL Quarterly, 40*, 439-451.

Neuman, W. L. (2000). *Social research methods* (4th ed.). Boston, MA: Allyn and Bacon.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research: A comprehensive guide*. San Francisco, CA: Jossey-Bass.

Smith, R. A., & Davis, S. (2003). The changing face of research methods. In S. F. Davis (Ed.). *Handbook of research methods in experimental psychology* (pp. 106-126). Malden, MA: Blackwell.

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage.

CHAPTER FOUR

EXPERIMENTAL RESEARCH DESIGN (EXD)

Conceiving and carrying out research is as much a creative process as it is a scientific one. (Seliger & Shohamy, 1989)

In this chapter you will learn the basic structure and function of experimental research, and the explanation and sources for further reading that will allow you to begin using this design in one of its many forms. You will also be introduced to key terms associated with EXD.

Experimental research design has a long and respected history in social science research, and those researchers working with this design have developed a rich vocabulary. One example is the term *variable,* since identifying and manipulating variables is important in this design. In other designs such manipulation is not desired. Many research articles do not explicitly identify the design they are using, therefore, recognizing the terms associated with specific designs becomes important to understanding.

Introduction

Experimental research design is usually regarded in one of three ways:

- It is the taken-for-granted research design
- It is considered the only possible design if one wants to be a serious researcher and establish a causal relationship. (This is not necessarily at odds with the first point.)
- It is an overused and increasingly irrelevant relic of the past with outmoded assumptions.

It may be the case that EXD is one of the most, perhaps *the* most, common research design currently used, and it is true that for many researchers the term *research design* is practically synonymous with the term *experimental research design*. No matter what you think about the experimental design, especially if you hold views one or three, it is helpful to become familiar with the experimental design in its various forms because all other designs tend to be compared, favorably or unfavorably, to it.

Experimental design defined

Cook and Campbell (1979) offer a description of experimental design that can be taken as a working definition: "All experiments involve at least a treatment, an outcome measure, units of assignment, and some comparison from which change can be inferred and hopefully attributed to the treatment" (p. 5).

- A *treatment* is something the researcher does. Often, language teachers want to evaluate the results of an innovation they have done in their class. In that case, the

innovation is the treatment.

- A *unit of assignment* is the persons or things the researcher studies.
- An *outcome measure* is typically a test that provides numerical data.
- A *true experiment* requires random assignment of participants to a control group as well as to another group, called either a treatment group or an experimental group.
- *Random assignment*, required to insure that both the control group and the experimental group are equivalent, is often impossible in educational research at the classroom level because TREES seldom control the assignment of students to classes and usually work with *intact classes*. Intact classes are classes assigned by administrative procedures or classes selected by students.
- When random assignment is not possible, the design is known as a *quasi-experimental design*.

What are the historical roots and key beliefs of EXD?

According to Ary, Jacobs, and Razavieh (1990), the roots of experimental design began in the 19th century with physical science. This was an attempt to improve observation by deliberately eliminating contextual complexity. It was found successful in scientific observation, and by the end of the 19th century, experimental design was applied to psychology and then to education. Experimental research design assumes variables can and should be identified and isolated from their context for study. EXD tends to rely on a Humean understanding of causality which maintains that causality *cannot* be seen, only inferred through a series of occurrences (Hume, 2004/1772). This is perhaps why users of EXD are drawn to inferential statistics to make comparisons.

Gass (2010) discusses the assumptions of EXD in two ways. First, second language researchers tend to be rationalists who believe that there is an objective reality; these researchers see a role for EXD. In other words, there are hidden assumptions and beliefs in EXD, and all designs for that matter. Second, Gass points to *inductive*-oriented research—which assumes data occurs before theory. This is as opposed to *deductive* oriented research—theory that occurs before data collection. She positions EXD in the deductive camp. The implications are that for confirmatory research (deductive), EXD is appropriate, but for research that is exploratory, EXD is probably not.

Key components of experimental design

Comparison is the essence of the experimental design; the EXD is predicated on the idea that the two groups being compared, for example control and experimental classes, are the same on all important aspects except the variable being investigated. Variables can be discussed within several categories:

- The *dependent variable* is the major variable that will be measured (Hatch & Lazaraton, 1991, p 63). A dependent variable cannot be identified in isolation because it is related to the independent variable. The dependent variable is the variable of focus--the central variable--on which the other variables will act. It is usually the test and what it represents. The dependent variable is the one we are trying to explain; there can be more

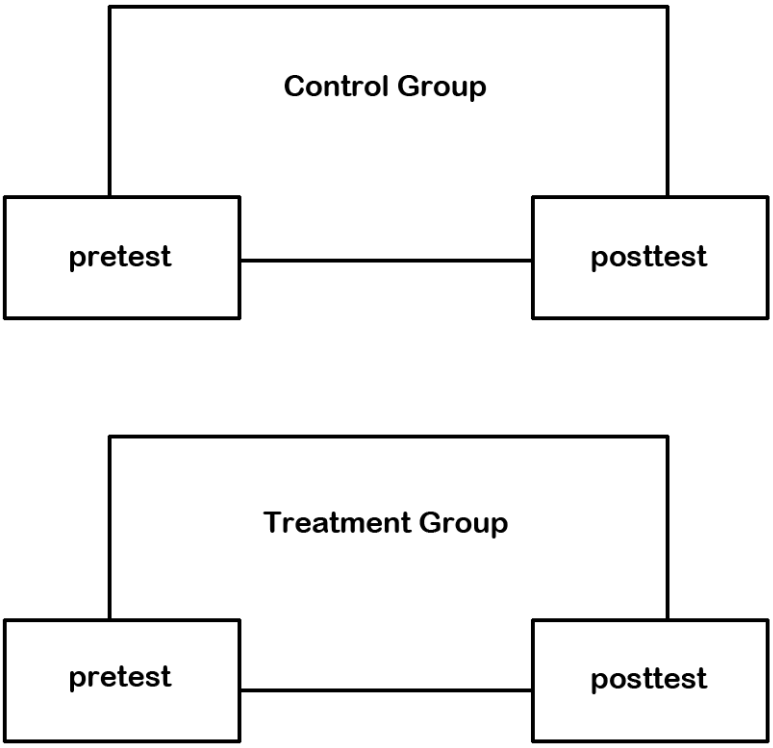
than one in a study.

- The *independent variable* in an EXD is the treatment; it is the variable that the researcher suspects may relate to or influence the dependent variable. In a sense, the dependent variable "depends" on the independent variable (Hatch & Lazaraton, 1991, p. 64). The researcher selects independent variables to determine their effect on or relationship with the dependent variables (Brown, 1988, p. 10).
- A *moderator variable* is an independent variable that the researcher does not consider important in the investigation (Hatch & Lazaraton, 1991, p. 65). In that sense, a moderator variable is a special type of independent variable (Brown, 1988, p. 11). A moderator variable is a question of degree. A moderator variable could affect the RQ, but would only modify, for example, it might be true for a special group (children but not adults). A moderator variable is treated statistically as an independent variable. Often a moderator variable is a surprise that the researcher finds later in the course of the research.
- A *control variable* is not of central concern in a particular research project, but might affect the outcome. *Control variables* are kept constant, neutralized, or otherwise eliminated so that they will not affect the study (Brown, 1988, p. 11).
- *Intervening variables* are abstract theoretical labels applied to the relationship that links the independent and dependent variables. They are constructs that may explain the relationship between independent and dependent variable, but are not directly observable themselves (Brown, 1988, p. 12). A variable that was not included in the study is the same thing as a moderating variable. The only difference is that the intervening variable has not been or cannot be identified in a precise way (Hatch & Lazaraton, 1991, p. 67).

The kind of data most closely associated with an experimental or quasi-experimental design are scores from a data gathering instrument, usually a test, that typically but not exclusively uses an interval or continuous scale. (For a discussion of scales, see Brown, 1988, p. 21). EXD is also closely associated with statistical analysis.

Comparison is facilitated by various methods of ensuring that the groups being compared are the same or similar; these methods include *random assignment* and sometimes *stratified sampling*. Random assignment ensures that students have an equal chance of being assigned to the control or treatment group. Stratified sampling, already discussed in Chapter Three, is a way to identify parts or strata of groups to be sure they are balanced.

Another important concept is *control*. Control means the elimination, or at least reduction of other factors except the ones under investigation. Control is important to insure that a comparison is valid because the experimenter is comparing what he or she thinks they are comparing.



How experimental research design might look as a visual

Figure 1. A visual representation of an Experimental Research Design

In Figure 1, a control group and a treatment group are compared at the beginning of the experiment by means of pretests, and are later compared at the end of the experiment by means of posttests.

What is EXD good at?

EXD is good at isolating and examining variables of interest: "The unique strength of experimentation is in describing the consequences attributable to deliberating varying a treatment (Shadish, Cook, & Campbell, 2002, p. 9). This is called causal description. EXD is also good at classroom research in situations where a TREE has a fairly high level of control over one or more classes. EXD is very good at describing groups and what happens in them. EXD can be used with small, intact groups, (see Hoyle, 1999) as well as a single subject (see Neuman & McCormick, 1995).

What is EXD not good at?

Despite the popularity of EXD, many points of concern and criticism exist:

1) Teachers with predetermined or intact classes cannot randomly assign students to a control or experimental group. Therefore, they must use a quasi-experimental design, a less powerful version of EXD.

2) Often the results that are compared are averages, and averages do not reveal what happened to individual students. For example, the results of an experimental group may not be statistically different from the results of a control group, but within the experimental group there may be clear cases of the curriculum working as well as equally clear examples of the curriculum not working, and an experimental design may make it difficult to examine these subgroups.

3) An experimental comparison does not in itself offer an explanation of what happened (Shadish, Cook, & Campbell, 2002, p. 9). A difference in scores between a control group and a treatment group that results in a low *p*-value and a high magnitude of effect does not explain the score difference. In other words, causal *description* does not equal causal *explanation.* Explaining *what* happened does not explain *why* it happened–for that, we need a theory that is external to an experimental design.

4) In order to make a valid comparison between groups, experimental design requires that other factors, called threats, do not intrude. With such attention being paid to threats, in order to establish internal validity, experimental and quasi-experimental designs may neglect external validity or generalizability.

5) Experimental design uses variables; in some research contexts, variables are not easily identified.

6) Many researchers claim that EXD is the only method that can claim causality: "If rival causes or explanations can be eliminated from a study then, it is argued, clear causality can be established" (Cohen, Manion, & Morrison, 2000, p. 211). Unfortunately, rival explanations can never be totally eliminated; one or more threats are always there. (See Shadish and Luellen (2006) for a discussion of threats.)

What are threats?

The concept of a threat is peculiar to experimental design because other designs, such as survey, case, and action research do not depend on the idea of variables and comparison. A threat, sometimes called an *alternative hypothesis*, is any possible explanation or reason for the results achieved other than or in addition to the one of interest. These other possible reasons are threats, so called because they threaten a claim.

It is helpful to think of a threat as "any condition which blinds or misleads researchers when they interpret their results" (Griffee, 2004). In other types of design, this condition may be called *bias*; in both cases, it reflects the same underlying problem of illusion. Bias may be introduced by instructor engagement with an innovation and a strong belief in its effectiveness. Threats are powerful because they rely on implicit--and thus hidden--beliefs and desires that can blind us to other realities, both internal and external.

Threats can often be anticipated, and to some extent, taken into account before research is initiated. This is preferable to having threats pointed out by an external reviewer after the research is complete, when it is impossible to change. Most of the threats discussed were first investigated and reported in Campbell and Stanley (1963), Campbell and Stanley (1966), Cook and Campbell (1979), and somewhat more recently in Campbell and Russo (1999).

Threat 1: History. History can be defined as "events, other than the experimental treatment, occurring between pretest and posttest and thus providing alternate explanations of effects" (Campbell & Russo, 1999, p. 80). To put it another way, things that are happening that the researcher is unaware of (Ary, Jacobs, & Razavieh, 1990).

One example of this threat includes ESL students taking an intensive course in an English-speaking country while making friends with native speakers outside of class, thus improving by participating in English conversations, regardless of the knowledge of English they are receiving in class (Long, 1984). Another example is the assessment of innovative teaching in basic writing courses. The researcher could be unaware of an outside source of writing improvement, For example, an increase in scores might be caused by some students visiting the university writing center, or taking additional classes with required writing assignments that contributed to writing improvement. Or, the instructor may be unaware that a student has a roommate or a friend who tutors her informally and helps her receive a high score as a result, which in turn raises the average score level. Any of these situations would constitute a threat of history to a research hypothesis that states the innovation under investigation caused the student writing improvement.

Controlling for the threat of history. The threat of history can be addressed by brainstorming possible effects external to the planned innovation. In the case of the writing course, the TREE can ask, "In addition to this innovation and my teaching, how else might my students be improving their writing outside of class, unbeknownst to me?" It is important, however, to judge the plausibility of the possible threats of history. Is it plausible that students are meeting secretly at night to discuss and revise their writing? Given what I know about this group of students, perhaps not. Is it plausible that some may be going to the campus writing center? Given that the writing center advertises their services, is on campus, accepts manuscripts over the Internet, and is a free service, it seems possible.

Two plausible threats were considered: (a) the possibility that students were going to the writing center and (b) that students were taking classes that required writing for which the instructor was actively helping them. The first threat was dealt with by assigning all students, both those in the control classes and those in the innovative curriculum classes, to go to the writing center. In fact, one meeting for all classes in this study was scheduled and held at the writing center. By assigning all students to go to the writing center, the threat that some students were going to the center was neutralized. The second threat, that some instructors were helping students, was dealt with by interviewing all students to see if that was the case. No such cases were reported.

When writing the results of a research report, threats of history should be identified and any actions taken should be described in detail. If space is limited, a brief mention may be made.

By thinking about the threats history and reporting thoughtful response to it, the researcher strengthens and increases validity.

Threat 2. The Hawthorne effect. The Hawthorne effect gets its name from a longitudinal study done in the Western Electric Hawthorne Works, Chicago (Roethlisberger & Dickson, 1939). It states that the mere knowledge that one is in a study may affect behavior. This is also known as the novelty effect (Beretta, 1992). Isaac and Michael (1995) describe causes of the Hawthorne effect as: novelty; awareness that one is a participant in an experiment; a modified environment involving observers; special procedures; new patterns of social interaction; and, knowledge of results in the form of daily productivity figures and other feedback, ordinarily not systematically available. If participants, probably students in our case, come to know they are in a special study, they may be impressed by the attention they are receiving, and this attention may cause them to do better, or at least act differently, then they normally would. This defeats the purpose of the experiment. Controlling for the Hawthorne effect increases both internal validity or causality and external validity or generalizability to other situations (Tuckman, 1995).

Controlling for the threat of the Hawthorne effect. Controlling for this threat entails conducting your study in an unobtrusive way so that participants are less aware they are being studied. This can be accomplished if you are the experimenter, your students are the participants, and your treatment is a normal part of your class curriculum. Alternatively, tell all participants, control and experimental alike, that they are subjects in a study. This is referred to as holding the knowledge constant or controlling for the threat. Controlling in this sense means holding constant or equalizing; the effect of controlling a threat is to neutralize it. A third possible way is to allow for an adaptation period during which no observations are made (Spector, 1981). You can also introduce a second control group that specifically controls for the Hawthorne effect (Tuckman, 1995). An irrelevant and unrelated intervention is deliberately introduced in order to create the Hawthorne effect so both groups have a Hawthorne effect. This is called the *placebo* effect. The *John Henry Effect* is the reverse of the Hawthorne effect. In this case, when subjects in the control group discover their status and, by that fact, are determined "to show the experimental group a thing or two," the control group sometimes outperforms experimental group (Isaac & Michael, 1995). *Resentful demoralization* (Lynch, 1996) is the opposite of the John Henry effect. In this situation the control group students make little or no effort because they believe they are in an inferior program.

Threat 3. Maturation. Maturation or maturing is the idea that participants may change over the time of the research, and this maturing affects the results, as opposed to the treatment. The change may be physical (age, fatigue) or psychological (interest or lack of interest). Long (1984) gives the example of ESL students in a U.S. program of several weeks who, because of residence in a native speaking country, may develop positive attitudes, which in turn increase motivation and thus achievement (that is, achievement unrelated to the program content).

Controlling for the maturation effect. One way to control for this effect is to select participants who are at the same developmental level and thus could be assumed to mature at about the same rate. A complete description of participants is helpful to readers.

Threat 4. Instability of data. This threat results from issues such as reliability, fluctuations in sampling persons or components, and instability of repeated or "equivalent" measures. In addition, instability may be a result of a large number of change-producing events, which taken individually are called *history* (Campbell & Russo, 1999). The smaller the population base, the greater the instability. One way of increasing data stability is to increase the number of participants. The instability of data threat is the only threat to which statistical tests of significance are relevant (Campbell & Russo, 1999).

Threat 5. Testing. This is also known as the *test effect* or *practice effect,* and refers to the effect on the scores of a subsequent test after taking an initial test. In other words, if participants take an initial test, learning may occur from taking the test, which affects the scores on the same or a similar test taken later on.

Controlling for the testing effect. To control for the testing effect, use counterbalancing in order to minimize the effect of having taken the same test previously. This is a cumbersome procedure in which two forms of the same test are used. On the first test occasion, some students take Form A while others take Form B. On the second occasion, the students who took Form A now take Form B, while the students who took Form B now take Form A (Brown, 1995). Another strategy is to use a control group that receives no treatment. If everything else is held constant, any practice effect that exists, and the size of the difference, should be reflected in pretest-posttest differences of the control group.

Threat 6. Reactivity. This occurs when the data collection instrument interacts with the treatment or even causes the treatment effect because it relies on self-reporting (Droitcour & Kovar, 2008). An example is a study asking participants to keep a log or diary of certain actions (say, using a dictionary), and the mere act of keeping the log makes participants more aware of the action (what words they looked up and how often they consulted the dictionary), thus causing them to either use the dictionary differently or more than they ordinarily would. If at the same time a treatment were involved, say direct instruction in dictionary use, and the participants increased their active vocabulary , was it because of the instruction, the reactive effect of the diary, or an interaction between them? Another example of reactivity (or test effect if the instrument is a test) is administering a questionnaire to a group of students asking them their opinion on a topic. Before completing the questionnaire, some of them may have had no opinion, but the questionnaire caused or tended to cause them to form opinions.

A third example of reactivity concerns social approval or disapproval of the activity being investigated. A researcher asks respondents if they get help on their English homework from native speakers. Some respondents may be reluctant to admit the true number of times they do or do not get help, and so distort their answers. Droitcour and Kovar (2008, p. 67) discuss this as *differential reactivity,* by which they mean there is a difference between the self-reported answers between those in the control group and those in the treatment group, which might happen because those in one group were treated differently. For example, in a study on smoking those in the treatment group were warned of the consequences of their behavior and those in the control group were not. Being warned of bad consequences may cause a group to act differently because of the warning and not because of what they normally do.

Controlling for reactivity In addition to self-reported data, the researcher could also include a form of data collection that did not involve self-report, say a form of teacher observation or collecting homework. For example, it might be possible to interview participants before the experiment to determine normal activity. When teaching in Japan, where students had a limited exposure to English, I conducted an experiment on English input during class. I surveyed the class to find out possible input sources, such as English-speaking friends, the number of English language movies they normally watched, and attendance at commercial language classes. After the experiment, I administered the survey again to determine if their activities had changed. By mentioning the survey and its results in my written report, I met the threats of reactivity and history.

Threat 7. Instrumentation. This threat concerns flaws in the testing instruments. One threat is that the instruments are not reliable or valid, or more likely, the researcher has not presented any reliability or validity evidence. A researcher is obliged to account for each and every data collection instrument used. This includes how reliability was calculated, what the coefficient was, how validation evidence was gathered, and what that evidence was. A second threat is that reliability and validity evidence was gathered in circumstance A, but the instrument is now being administered in circumstance B. For example, a listening test is created for the purpose of testing ESL students in North American high schools, but is administered to first-year university EFL students in Japan. The circumstances are different, so the validation evidence from North America does not apply to the circumstances in Japan.

A third form of instrument threat involves the impact of variations in the test. For example, if version A of a test was used in the pretest, but version B was used on the posttest, any differences in performance could be due to discrepancies in the versions of the test (the instrument) rather than the treatment (Long, 1984). A fourth cause of instrument threat refers to a "shifting of the measurement instrument independent of any change in the phenomenon measured" (Campbell & Russo, 1999, p. 84). In other words, the data collection instrument can change over time. One way this might happen is when record keeping is centralized or in other ways changed. For ESL researchers, it might be using different forms of the TOEFL. Or it might be using a test for one group and using an interview for another group, and comparing the groups using two types of data. Another way a test could change over time is with interviewing. Interviewers may improve their techniques with practice and later interviews may be more valid than earlier interviews due to improved skills (Spector, 1981, p 26).

Controlling for the threat of instrumentation. To meet the first and second forms of the threat, namely test validity, the researcher can pilot the test instrument under local conditions and report reliability and validation results. For the third form of the threat, the researcher can use the same test as a pretest and posttest, if it is judged that there is adequate time between administrations so that the testing effect is not an issue. If two versions of the same test must be used, the researcher can pilot the tests and report data showing the test results are statistically not different. To address the fourth threat, namely change in the test instrument, the researcher can verify that the test did not change.

Threat 8. Regression. Also known as statistical regression, this can be defined as "[p]seudo-shifts occurring when persons or treatment units have been selected on the basis of their extreme scores" (Campbell & Russo, 1999, p. 80). It can occur when participants have been selected on the basis of extreme scores (Isaac & Michael, 1995). It is likely to happen in remedial courses when groups are selected on the basis of extreme low scores (subjects failed a proficiency test) or with groups having unusually high scores (advanced classes of mostly high proficiency students). It could also happen in a class when the TREE administers a pretest on a day when many persons are sick. Scores might be artificially depressed, and rise over the period because students became healthy. Regression might also be an issue if some of the more proficient students are absent on the pretest, but present to take the posttest. The threat is that the TREE believes scores have risen due to the treatment when, at least to some extent, they may have risen due to the natural tendency of group scores to rise over time.

Threat 9. Selection. Selection occurs when the TREE forms the comparison groups. The problem may be that the groups are different to begin with, and as a result produce different scores. The issue becomes one of comparing apples and oranges. In other words, are the groups really equivalent? The threat is that it is this difference in groups that produces the score changes, not the treatment. Long (1984) says that students selected to be in one group may differ in some important way from students in the other groups, for example, they may be more motivated or more intelligent. For example, if one group comprises volunteers and another group is required to take the course, selection bias would be suspected. Similarly, if the students to be evaluated are from a particular socioeconomic group, this may explain the results--independent of the program--especially if the other students are selected based on other criteria (Lynch, 1996).

Controlling for the threat of selection. When random assignment is not possible, a researcher should attempt to control differences among groups as much as possible. Matching participants or variables within the groups (for example, personal characteristics and demographics), can achieve this. If pretests are administered, another strategy is to compare the scores between the groups. If the descriptive statistics indicate that the scores are similar, the TREE can argue that the groups are not different in any significant way. Class grades can also be used in the same way.

Threat 10. Mortality. Mortality refers to the loss of students in either or both the control or experimental groups. Students often drop out of a program, sometimes in large numbers. When students drop out of a program, it may create a threat that requires investigation. Perhaps several students drop out of the experimental group, leaving only the better students, and their scores are higher, inaccurately indicating that the program was better.

Controlling for the threat of mortality. If the dropout rate for a program is considered high, good recordkeeping may be useful. The researcher should keep a record of students who drop out, including their contact information. It may be possible to contact them and interview them as to their motivation. Any test scores or grade indicators should also be saved because it may be possible to show the achievement level of the students.

Threat 11. Researcher expectancy. When a researcher expects certain outcomes, and thereby causes them to occur (Brown, 1995), this is referred to as researcher expectancy. It is unlikely

that a TREE would intentionally cause favorable results, but threats deal with illusion, which can be powerful. There may be many ways a TREE can cause results to be favorable to the desired outcome, especially if the grading is not objective.

Controlling for the threat of researcher expectancy. TREES using qualitative data collection are familiar with this problem, and recommend keeping a research log in which to write expected results. Such a record can alert a TREE to this threat because if a researcher expects something to happen, and it happens just as expected, this could serve as a warning to be especially careful and critical. Another control is to create an explicit grading criteria, and ask colleagues to grade at least some tests and homework assignments. Correlating scores against colleagues' is a good way to see if grading is being done in a consistent manner, especially if the researcher is grading work in both the control and experimental groups.

Threat 12. Teacher effect. The teacher effect is not a threat if a TREE is conducting an experiment and is teaching both the control and experimental classes. If this is not the case, the teacher effect can occur when a researcher is using multiple teachers, say one in the control class and another teacher for the experimental class. The threat is that one teacher may be more proficient than another, and it is this proficiency that accounts for the change or lack of change in the experiment.

Controlling for the threat of teacher effect. One strategy is to assign teachers randomly to treatment and control groups from a large pool. Another strategy is to have both groups taught by the same teacher(s). One could also standardize the teaching, and then observe classes to see if the standardized teaching was accomplished. Another strategy is to eliminate teachers from the experiment by using a recording of the class. Finally, the researcher may have the option of examining processes in the classroom by researcher observation.

Threat 13. Diffusion of treatments. This refers to a situation when the control group is given the treatment intended for the experimental group. This is typically done by a teacher who changes the regular course of instruction, believing the innovation to be helpful (Lynch, 1996).

Controlling for the threat of diffusion of treatments. One remedy is to select a teacher who understands research and will adhere to the teaching protocol. Another strategy is to not let the other teacher know what is being done in the other class. A third possible solution is asking the other teacher for support by explaining that it is not known whether the experimental treatment is better or not (which is, of course, true because if it were known, there would be no need for the research), and that research is needed to determine this. A fourth solution is to arrange for the control group teachers to explain in detail what they plan to do and why they think it is an effective teaching curriculum. Teachers don't change the curriculum if they are convinced it is effective. A fifth solution is to create an alterative curriculum for the control class that has instructional value.

Threat 14. Ecological representativeness. This is an external threat to validity, meaning that it is a threat or hindrance to the ability of results to be applicable or generalizable to other classrooms or situations. Most of the threats mentioned so far are threats to internal validity in that they hinder the internal argument for causality. According to Beretta (1986), "The more

the setting of our evaluations resemble regular classrooms, the greater the degree of ecological representativeness and the more confident we can be in extrapolating to other settings" (p. 147). By contrast, a study that takes place under tightly controlled conditions has no credible relationship with what might happen in actual classrooms. This is one of two reasons TREEs don't do more pure experiments; the other is that we are usually given intact classes.

Many TREEs, after reading and thinking about the multiple threats discussed here, begin to wonder how they or anybody could possibly do any research, especially that using experimental design. If you happen to encounter a colleague who is depressed about the many threats that haunt research, here are some comments you can use to cheer them. First, not all threats apply to every research plan. In fact some threats are opposites. For example, it is not possible to experience both teacher effect and researcher expectancy simultaneously. It is not even likely that most threats need to be addressed in one research project. A TREE should consider all the threats discussed here to decide which are most likely to be applicable. The point is to be on alert for threats, not to be depressed by what might go wrong. Second, a threat is just that, a threat, not an actuality. Researchers need to convince readers they have been diligent, not that the research is perfect. Third, researchers consider threats so that problems can be anticipated, steps can be taken against them, and evidence can be gathered to show that this was done in the course of the research. This is obviously preferable to having a problem pointed out after the class is over, the students have dispersed, and there is little or nothing that can be done to correct the problem.

Trends in Experimental Design

There are recent indications that EXD, and data from various types of tests with which EXD has had a close connection, are being reappraised. Several researchers suggest that those who use EXD are sensitive to the influences of emergent methods, qualitative research, and postmodern thinking, and that proponents of EXD are responding to those influences.

Mark (2008), for example, offers several trends, two of which are reviewed here. First, grandiose claims for EXD research are replaced with more modest claims. Instead of claiming that experimental design is the "gold standard" and the one and only way to conduct rigorous research, EXD is described as appropriate under some circumstances and not appropriate under others. For example, EXD is appropriate when effects of a variable are desired (say, the effect of a certain teaching technique on test scores), but not in other circumstances such as when the cause of an effect is desired (for example, the scores on proficiency tests are decreasing and the cause is unknown).

A second trend is the interactive and emergent aspect of EXD. One example is the process many researchers who use experimental design go through when they anticipate threats, and what they do if an unanticipated threat materializes after the research. The emergent process of actual research using EXD is often hidden by a style of reporting that begins with a clearly defined problem, proceeds with a logical methodology, and concludes with well presented conclusion. This reporting style hides the twists and turns that the EXD researcher actually went through. The trend is not that suddenly EXD has become emergent, but that it is now possible to state the emergent nature of EXD that always existed.

Frances Stage (2007) edited a short book (95 pages) comprising eight articles by researchers on the relationship of quantitative research and critical questions. Critical research is characterized by asking questions about the power relationships in society and how previously dispossessed or neglected groups can be transformed by research. Stage (2007) describes this collection on articles by saying:

> This volume describes an increasingly evident way of conducting quantitative research. This approach does not seek merely to verify models. Rather it focuses on questioning and then modifying models or creating new models that better describe the ever-differentiating individuals who are the focus of educational research. (p. 9)

The focus of research moves away from the technical problems to the persons in society who might be affected by research, and how their causes might be identified and assisted.

The theme of focusing on society is continued by Padilla and Borsato (2008) in the third edition of the *Handbook of Multicultural Assessment*. They, as well as Malgady and Colon-Malgady (2008), concentrate their comments on norm-referenced tests and the all-important norm group, so frequently used in experimental research design. These writers emphasize that the norm group must represent the various groups tested. Of course, the idea that the norm group must represent the group being tested, or vice versa, is not new. However, Padilla and Borsato (2008) emphasize that test scores must be interpreted, and Malgady and Colon-Malgady (2008) emphasize that for tests to be fair to diverse groups as defined by gender, sexual orientation, religion, or demographics, the norm group must be sensitive to those groups.

The larger trend in modern research practices is towards adapting traditional research designs, such as the experimental design, and traditional data collection instruments, such as tests, to the multitudinous and diverse concerns of today. Practioners of these designs and instruments are responding--evidence of the underlying vitality and flexibility of the research enterprise.

Three examples

The following three examples of typical experimental research designs describe single class, two classes, and two classes in a special adaption of the EXD called time-series design.

Single class. In the first type of EXD, one class is measured by a test (the dependent variable) purporting or claiming to measure some category (the independent variable) at the beginning of the term. Some sort of treatment or innovation is done, and the class is measured by the same or similar test at the end of the term. Scores from the pretest are compared to the scores from the posttest; an increase in the posttest scores is taken to mean that the treatment or innovation succeeded.

Two classes. In the second variation of an EXD design, two classes are measured, as seen in Figure 1. One class is the control, in which the traditional or usual way of teaching is done, and the other class is the experimental class, in which some innovation is introduced. The pretests are administered to both classes; similar scores are used to argue that the classes are not

different, and thus a comparison is valid. Any data can be used to establish initial group equality: for example, gender mix, age distribution, proficiency level, or ethnicity. The scores from the posttest instrument are then compared, usually through one of three ways.

The most common way of comparing two groups is through an inferential statistic significance test such as a *t*-test (see Appendix C). If the treatment group scores are statistically significantly higher, it is taken to mean the innovation succeeded. In other words, the goal is to demonstrate that the two sets of pretest scores show *little* difference and that the posttest scores show a *big* difference. There are many variations of this pattern, and the TREE has to study these design variations to see which best applies to his or her situation. For a detailed checklist, see Creswell (2003, p. 163). A second way of comparing two groups (or more, but two is the most common) is to use non-parametric statistics. They work essentially the same as parametric statistics, but are used when for some reason (low N-size or non-normal distribution) normal distribution is not achieved. A third way of comparing groups is graphic design. Figure 2 is an example.

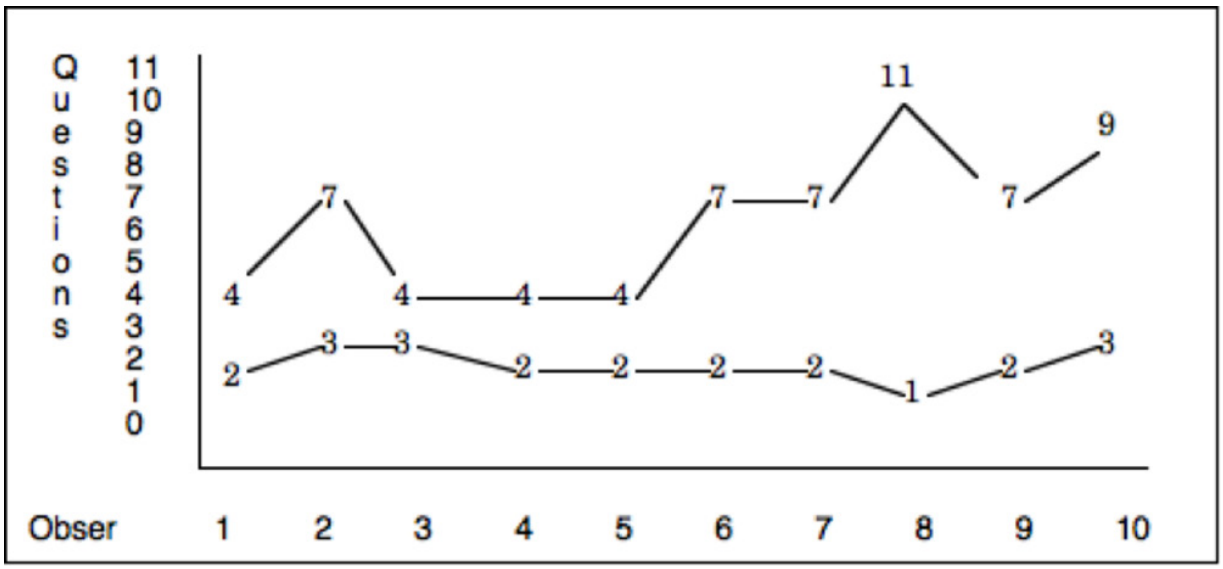


Figure 2. Results of the two classes showing number of questions students asked.

Note: Obser = observation occasions, top line = number of questions asked by the experimental class, bottom line = the control class.

Time-series. The third example is a time-series design from an unpublished study. The purpose of the study was to document the effectiveness of clarification response training. I wanted to investigate to what extent the explicit teaching of a clarification model increased both the number and quality of student questions. One of my research questions was: Will the number of student questions to the teacher about the meaning of vocabulary words increase after the explicit teaching of a clarification model?

There were a total of 38 student participants in two second-year intact classes of Academic

English at a small, liberal arts university near Tokyo, Japan. One class of 20 students (15 men and 5 women) met on Monday and Wednesday, and the other class of 18 students (10 men and 8 women) met on Tuesday and Thursday. The Monday class was randomly chosen to be the experimental group, and the Tuesday class was used as a control group. Both classes met for 90 minutes two mornings a week and had the same teacher. All but one or two students in both classes were 19 or 20 years old, and most students from both classes were from the local area.

Materials included the clarification model, a student language background data information questionnaire, a threat of history questionnaire, and vocabulary worksheets not provided. Both classes were taught in the same classroom.

The classes were told that I was interested in recording their questions and comments about vocabulary. The innovation or treatment was a clarification model, which promoted asking for clarity and paraphrasing. It was introduced to the Monday class on the twelfth class meeting, which was the class before the sixth vocabulary worksheet was handed out. For the control class, instead of question training, fluency training was substituted (Griffee, 1994, p. 8). Only the portion of the class dealing with the vocabulary questions was recorded. The average time for this part of the lesson was typically five to ten minutes.

The audiotapes were transcribed. Two native speaker teacher colleagues, one male and one female, were selected and trained as raters. Training consisted of an explanation of the purpose of the research, definitions and examples of number and quality, and a sample transcript. Each rater judged the number and quality of questions on the sample transcript, and discussed the results with me. Raters were given the ten class transcripts and asked to decide both number and quality of student questions. When raters disagreed, the researcher arbitrated.

A time-series design (a variation of an experimental design) was used to record the extent to which the number and quality of student questions increased after the clarification model was introduced. Time-series research entails several observations being taken to establish baseline data, a treatment being introduced, and observations continuing to determine whether the treatment had an effect (Ary, Jacob & Razavieh, 1990). Rossi, Freeman, and Lipsey (1999) characterize time-series as a large number of repeated measures (at least 30) on a relatively large population. Statistical tests are then used to determine a projected trend. They contrast time-series with panel studies, which use a relatively modest number of repeated measures on the same group for the purposes of studying the effects of a program. However, for most researchers, time-series is the generic term. Accordingly, I will use the term *time-series* to refer to a number of repeated observations on the same group.

One advantage of a time-series design is that an intact class can function as its own control group (Hatch & Lazaraton, 1991). Another advantage is that threats to validity, such as the Hawthorne effect, maturation, regression, and test effect can be controlled for (Ary, Jacob & Razavieh, 1990; Mellow, Reeder & Forster, 1996). A time-series design assesses these threats because frequent observations before the introduction of the treatment allow for the collection of baseline data, which provides evidence for these threats. The major threat to a time-series design is history (Spector, 1981, p. 31). As Cook and Campbell (1979, p. 211) point out, history is the possibility

that events other than the treatment influenced the dependent variable (the number and quality of questions) immediately after the treatment was introduced. Because my data comprised frequency data and because of the relatively low number of interventions (5 pre-interventions and 5 post-interventions), I followed Mohr (1988, p. 150), who suggests that one good way to analyze trends in a time-series design is visual analysis of a raw data graph. Inter-rater reliability was calculated by Spearman correlation using StatView statistical package for the Macintosh (StatView 5.0, 1998). Qualitative data were gathered in a teacher log I kept for both classes. During or immediately after each class, I recorded environmental issues such as attendance, notes on individual student performance, and my reflections on the research process.

Interrater reliability as calculated by Spearman correlation was .98 ($p < .003$) for the Monday class and .89 ($p < .008$) for the Tuesday class. Figure 2 shows the results. The top line is the Monday class (the experimental class), and the bottom line is the Tuesday class (the control class). Treatment began after the fifth observation. Except for observation two, the average number of questions asked in the experimental group is about four and the average number of questions for the control group is about two. After the introduction of the question model starting from vocabulary worksheet six, the control group continues to ask about two questions.

Internal validity concerns the issue of causality. In other words, did the clarification question model cause the increase of questions in the experimental class, or might there be another cause? Maturation, test effect, and regression are dispelled by the use of a time-series design and a control class. The chart in Figure 2 shows a fairly flat response over the first five observations for both classes, and then an increase for the experimental class, but not the control group. If maturation, test effect, or regression were the cause of the increase in student questions, the line on the chart would be gradually sloping upwards. The threat of researcher expectancy, that is the researcher finding what he wanted to find, was dealt with by using two outside raters to count both question number and quality.

The Hawthorne Effect was countered by the time-series design. In addition, before collecting baseline data, I told both groups that I was interested in their questions about vocabulary. The design shows no change in the experimental group until the question model was introduced, and no change at all in the control group. If the students in the experimental class were acting in a way that was influenced by the knowledge that they were involved in a project, why would there be a sudden increase in the number of questions only after I taught the question model? In addition, since I told both groups about the project before the first observation, and both classes saw the tape recorder on the table for all ten observations, it is not likely that they suddenly realized they were being observed, and this knowledge caused a change in class performance.

The threat of history was taken into account by use of both a control group and questionnaire. The use of a control class was helpful because any external event that might have caused an increase in questions in the experimental class would also have been noticed by the control group and caused a corresponding increase. A questionnaire was used to see if there were any sudden increase in English language input at the time of the question model introduction. The questionnaire was administered early in the semester to determine how much English language input students received outside of class from viewing English language movies, videos, TV,

radio, friends, home study, or study at a commercial language school. The questionnaire was re-administered toward the end of the semester to 18 students in the Monday class (3 absent) and 17 students in the control class (1 absent). The second administration was given to determine notable changes in language input at the time I introduced the question model. There were no changes in the experimental class, and only slight changes in the control class (2 students in the control class reported an increase of one or two more movies and videos). I conclude from the questionnaire results that none of the students in the experimental class experienced any increase in English language that would account for the increase in questions input after the introduction of the question model.

Further Reading

Creswell (2003) discusses experimental design. Two of the classic sources for experimental and quasi-experimental design are Campbell and Stanley (1963) and Cook and Campbell (1979), but the most current and easily available text today is Shadish, Cook, and Campbell, (2002). Heppner, Kivlighan, and Wampold (1999) have an extensive discussion of experimental, quasi-experimental, and time-series designs. Hatch and Lazaraton (1991, p. 84) discuss large N-size designs (30 participants or more) one-shot design; one-group, pretest-posttest design; intact groups, single control, pretest-posttest design; and time-series design, and Hoyle (1999), Neuman and McCormick (1995), and Saville and Buskist (2003) discuss small N-size (10 participants or less) also known as single-subject experimental research. Examples of studies using the experimental research design can be found in many journals. One example is Folse (2006) who uses a modified EXD to compare three groups of vocabulary learners using three types of written exercises.

DISCUSSION QUESTIONS

Write any questions you had while reading about experimental research design.

Reflection on Experimental Research Design

1. What is the attraction of Experimental design for classroom teachers in general and/or you in particular?

_____

_____

2. What problems or issues would you anticipate in using EXD?

_____

_____

Task 1

Find a published research article that you believe used an experimental, quasi-experimental, or time-series design. Download, copy, or print the article, and answer the following questions. Bring the article and answers to class.

1. State the title of the article, the author, year, and the name of the journal.

2. What do you find in the article that tells you the design was experimental or quasi-experimental?

3. Where did the research take place, and who were the students?

4. Can you identify the dependent and independent variables?

5. Were any threats mentioned? How were they controlled?

6. Were you convinced by the research? Do you accept the conclusions of the author(s)?

Task 2

1. What innovation have you done are you thinking of doing in your class? If you are not teaching a class, think of a change or innovation that you have tried or might be interested in trying in the future.

_____

_____

2. What would you compare and how would you compare it?

_____

_____

3. Can you think of a way to get a control group?

_____

_____

4. What threats would be problematic?

_____

_____

Glossary of Key Experimental Design Terms

Control group A group of participants used for comparison purposes, similar to the experimental group, but to whom no treatment is administered. What happens in the control group, it is argued, is what would happen if no experiment or treatment were done. Results from the control group are known as *counter-factual evidence*.

Dependent variable A variable that is usually the test used to measure results. In terms of cause and effect, this is a measure of the effect.

Experimental group: Also known as the treatment group, the group of participants with whom the innovation or treatment is applied. The experimental group is where change of some sort will be compared with the performance in the control group.

Hypothesis A hypothesis is a possible explanation. The *null-hypothesis* states that there is no relationship between variables of interest. The *research hypothesis* is the preferred explanation put forward by the TREE to explain the results of the experiment. An alternative or *rival hypothesis* is an explanation other than the research hypothesis that also might explain the results.

Independent variable The variable under investigation. For example, if a TREE were using EXD to investigate listening comprehension, then listening comprehension would be the independent variable. If a TREE were investigating how students acquire vocabulary, then vocabulary acquisition would be the independent variable.

Inferential statistics A class of statistics that does not directly describe scores, but allows the TREE to infer or suppose something that cannot be directly observed, usually the probability that two scores are related (correlation) or not related (a significance test).

Intact class A class that is whole or intact before the TREE comes into contact with it. A class is intact because either it forms itself (students select the class and then enroll for some reason such as convenient time) or the class is formed by others (administrative control). Intact classes are the norm because most teachers cannot randomly assigned students.

Mean scores The average score for a group of taking a test. A mean score of 82.7 indicates that the average score for the whole class was about 83 points.

Non-parametric statistics A family of statistical procedures used when the requirements of parametric statistics are not available. Most often this is because the scores do not form a normal distribution or the test does not use continuous scale. A popular non-parametric statistic is the Chi-square, which uses frequencies.

***p*-value** The small letter *p* stands for probability. A TREE administers a test to two groups of students and wonders if the difference between the resulting scores is likely the result of chance variation. The answer comes in the form of a percent called the *p*-value. If the *p*-value is small, traditionally .05 or smaller, the TREE can make the claim there is a small probability, 5%, that they would get these scores if there were no relationship; therefore, they can *infer* that there is a relationship. If the point of the experiment was to show score improvement, the TREE hopes the

mean difference is large and the *p*-value is small.

Parametric statistics A family of statistical procedures that requires normal distribution and a continuous scale. Some familiar parametric statistics are the t-test, ANOVA, and Pearson correlation.

Pretest A test that is administered before the experiment, intervention, treatment, or teaching takes place. The purpose of the pretest is to establish baseline data; in other words, it establishes what the situation is before the intervention or treatment is administered.

Posttest A test that is the same or very similar to the pretest, and is administered after the experiment or teaching takes place.

Random assignment A type of placement that refers to the process by which any participant has an equal chance to be assigned to any group. One way of random assignment is to write names on pieces of paper, put them into a container, and have somebody draw them out one by one without seeing the names. Another way is to use a computer program that can randomize a list of names or numbers.

Reliability The idea of consistency of results or getting the same or similar results each time data is collected. It was originated by Spearman as a comparison of a hypothesized true score and observed score. It is often given in terms of a percent marked with a decimal point, such as .82. See Appendix A on Reliability.

Scale Refers to unit of measurement. Every measuring device must have a scale in which the measurement is reported. Tests are usually reported in one of three scales: nominal (e.g., gender, nationality), ordinal (e.g., frequency counts), or continuous (e.g., test scores).

Significance test A statistical procedure that estimates the difference between two (sometimes more depending on the type of test) sets of scores. The results of a significance test is a *p*-value, which is an indicator of the likelihood of obtaining this difference given the null hypothesis. A common significance test for testing the difference between two and only two scores is called a *t*-test. Another significance test is called analysis of variance (ANOVA) and is used for testing the difference between two or more groups. See Appendix C on Significance Testing for a fuller discussion.

Stratified sampling Although identified with SRD, stratified sampling can be used in EXD to equalize groups. For example, a TREE considers ethnic groups important variable. The TREE has a control group with 50% one ethnic group and 50% another group. In order to make the treatment group similar to the control group, the TREE wishes to use only those students in the two ethnic groups in approximately a 50-50 proportion.

Threats The name given to a rival hypothesis used to explain the difference between treatment groups. For example, if a researcher were investigating the effects of an innovative writing curriculum, a threat would be that, unknown to the researcher, many students went to the writing center, and those visits accounted for the score increase, not the innovative curriculum.

Attendance at the writing center, usually called a threat of history, threatens the research hypothesis that it was the innovative writing curriculum that caused the increase in scores.

Time-Series A variation of EXD where instead of comparison of two groups, the TREE uses only one group, takes several measurements to establish baseline data (to establish which is usually the case), introduces an innovation or treatment, and continues with additional measurements (to establish the effect of the treatment). The main advantage is only one group is necessary and can function as its own control group.

Unit of Assignment The persons or things being studied. Things being studied can be classes, schools, or even cities and towns. An EXD can study one person, a group of persons, an entire class, or a school.

Validity Test validity is traditionally defined as providing evidence that shows or indicates that what the TREE believes he or she is testing is actually being tested. In practice, validity means offering an explanation of how one checked or verified instrument results.

Variable A trait or quality that can be measured by a test or other kind of data collection instrument. A variable can be anything a TREE is interested in.

References for Experimental Research Design

Ary, D., Jacobs, L., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Orlando, FL: Harcourt Brace.

Beretta, A. (1986). Toward a methodology of ESL program evaluation. *TESOL Quarterly, 20*(1), 144-55.

Beretta, A. (1992). Evaluation of language education: An overview. In J. C. Alderson & A. Beretta (Eds.). *Evaluating second language education* (pp. 5-24). Cambridge: Cambridge University Press.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (1995). Language program evaluations: Decisions, problems and solutions. *Annual Review of Applied Linguistics, 15*, 227-248.

Brown, J. D. (1997). Designing a language study. In D. T. Griffee & D. Nunan (Eds.). *Classroom teachers and classroom research*. Tokyo: Japan Association for Language Teaching.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.). *Handbook of research on teaching* (pp. 171-246). Chicago, IL: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Campbell, D. T., & Russo, M. J. (1999). *Social experimentation*. Thousand Oaks, CA: Sage.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Droitcour, J. A., & Kovar, M. G. (2008). Multiple threats to the validity of randomized studies. In N. Smith & P. Brandon (Eds.), *Fundamental issues in evaluation*. New York, NY: Guilford Press.

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly 40*, 273-293.

Gass, S. (2010). Experimental research. In B. Paltridge & A. Phakita (Eds.), *Continuum companion to research methods in Applied Linguistics*. London: Continuum Publishing.

Griffee, D. T. (1994). Circle conversation. In K. M. Bailey & L. Savage (Eds.). *New ways in teaching*

speaking (pp. 8-10). Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Griffee, D. T. (2004). Research tips: Validity and history. *Journal of Developmental Education, 28* (1), 38.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Newbury House.

Heppner, P. P., Kivlighan, D. M. Jr., & Wampold, B. E. (1999). *Research design in counseling* (2nd ed.). Belmont, CA: Wadsworth.

Hume, D. (1772/ 2004). *An enquiry concerning human understanding*. New York, NY: Barnes & Noble.

Hoyle, R. H. (1999). *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage.

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation* (3rd ed.). San Diego, CA: Educational and Industrial Testing Services.

Kelly, K. (1994). *Out of control: The rise of neo-biological civilization*. Reading, MA: Addison-Wesley.

Long, M. (1984). Process and product in ESL program evaluation. *TESOL Quarterly, 18*(3), 409-425.

Lynch, B. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University Press.

Malgady, R. G., & Colon-Malgady, G. (2008). Building community test norms: Considerations for ethnic minority populations. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed.) (pp. 34-51). San Francisco, CA: Jossey-Bass.

Mark, M. M. (2008). Emergence in and from quasi-experimental design and analysis. In S. N. Hesse-Biber & P. Leavy (Eds.), *Handbook of emergent methods* (pp. 87-108). New York, NY: Guilford Press.

Mellow, J., Reeder, K., & Forster, E. (1996). Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition, 18*, 325-350.

Mohr, L. B. (1988). *Impact analysis for program evaluation*. Chicago, IL: Dorsey Press.

Neuman, S. B., & McCormick, S. (1995). *Single-subject experimental research: Applications for literacy*. Newark, DE: International Reading Association.

Padilla, A. M., & Borsato, G. N. (2008). Issues in culturally appropriate psychoeducational assessment. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment:*

Clinical, psychological, and educational applications (3rd ed.) (pp. 5-21). San Francisco, CA: Jossey-Bass.

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker: An account of a research program conducted by the Western Electronic Company*. Cambridge, MA: Harvard University Press.

Rossi, P. H., Freeman, H. W., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Saville, B. K., & Buskist, W. (2003). Traditional idiographic approaches: Small-N designs. In S. F. Davis (Ed.). *Handbook of research methods in experimental psychology* (pp. 66-82). Oxford: Blackwell.

Seliger, H. W. & Shohamy, E. (1989). *Second language research methods.* Oxford: Oxford University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., & Luellen, J. K. (2006). Quasi-experimental design. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 539-550). Mahwah, NJ: Erlbaum.

Spector, P. E. (1981). *Research designs*. Newbury Park, CA: Sage.

Stage, F. K. (2007). (Ed.). *Using quantitative data to answer critical questions*. San Francisco, CA: Jossey-Bass.

Stage, F. K. (2007). Answering critical questions using quantitative data. In F. K. Stage (Ed.), *Using quantitative data to answer critical questions* (pp. 5-16). San Francisco, CA: Jossey-Bass.

StatView 5.0 [Computer software]. (1998). Cary, NC: SAS Institute.

Tuckman, B. (1995). *Conducting educational research* (4th ed.). New York, NY: Harcourt Brace.

CHAPTER FIVE

CASE STUDY DESIGN (CSD)

Case study research is a strategy for doing social inquiry, although what constitutes the strategy is a matter of some debate. (Schwandt, 2007, p. 28)

In this chapter you will learn the short but interesting history of case study design (CSD), that case study (like experimental design) generalizes to theory, not population, and how you might implement a case study.

Preview Questions

1. Can you think of some examples of case histories?

2. Have you ever read a case history?

Introduction

For many, case study design is considered naturalistic and qualitative. For example, Brown and Rogers (2002), Creswell (2002), and McKay (2006) discuss CSD in terms of qualitative data in the ethnographic tradition. In fact, it is not uncommon to divide all research into experimental and case study approaches. This identification of CSD with qualitative research may be the reason that CSD is experiencing a comeback. Despite, or perhaps because of its increasing popularity, the use of CSD is prone to confusion because it is not well defined. Much of what is called CSD resembles administrative reports, for example, "This is my program and this is how it works" (see Kaufman & Brownworth, 2006).

Case Study Design Defined

Case study is occasionally confused with experimental single-case design--an experiment using only one group. Sometimes CSD is confused with time-series experimental design, another instance of an experiment using only one group (Nunan, 1992). "A case study or ethnographic research project may seek to answer specific questions about occurrences and their explanations similar to those answered by quantitative" oriented researchers (Tuckman, 1999, p. 401). In addition, case study design is also confused with ethnography; it is easy to see why. Babbie (2004) defines ethnography: "An ethnography is a study that focuses on detailed and accurate description rather than explanation" (p. 289). That definition could also fit a case study, and suggests that at some level a case study and an ethnographic study might overlap to the point where they would be the same. This is exactly what Nunan (1992) says: "I would agree that the case study resembles ethnography in its philosophy, methods, and concern for studying phenomena in context" (p. 75). However, he goes on to suggest at least three ways a case study and an ethnographic study might differ. One way is that ethnography covers a larger scope; CSD is more limited and more focused. Second, an ethnography is more likely to attempt to define the culture included in the study, while CSD is more likely to investigate narrowly defined topic areas

such as classroom problems or language development. The third way Nunan (1992) posits a possible difference is that while both use qualitative methods, CSD can also use quantitative data collection methods. Hamel (1993) suggests that "a case is an in-depth study of the cases under consideration" (p. 1).

What Is a Case Study?

For Yin (2000), a case study must have three aspects, which can serve as a definition: it must have data from multiple sources, examine something in a real-life context, and use theory to generalize results. Yin (2003) elaborates this definition by saying that a case study design is used when the difference between the object of the study and the context of the study cannot be easily differentiated.

What are its historical roots and assumptions?

Case study has a long history in anthropology dating from Bronislaw Malinowski's work in early 20th century Melanesia. According to Hamel (1993), Malinowski considered the isolated tribe or village an ideal unit in which to study culture. Case study entered American sociology by means of the sociology department of the University of Chicago, where Robert Park, a former journalist, developed research methods for direct investigation, including open interviews and material collection (Hamel, 1993). In the 1930s, a political and methodological argument developed between forces in New York (at Columbia University) and Chicago (University of Chicago). Columbia championed the statistical survey while Chicago accepted both the statistical survey as well as the case study. By the end of the 1930s, the statistical survey and statistical methods had won and the case study and fieldwork had lost (Hamel, 1993). This now forgotten battle probably accounts for the less highly regarded status that many researchers feel for the case study, but as qualitative research experiences resurgence, CSD is being reconsidered.

The beliefs and assumptions of CSD maintain that a large group, say a culture or society, can be understood by studying a smaller unit of that society. For Malinowski, the key unit was an isolated tribe or village in which culture could be studied in isolation, free from the influences of European culture. For the French anthropologist Le Play it was the French working class family, and for members of the Chicago School of sociology it was immigrants newly arrived in Chicago. In each case (pun intended), the smaller unit was considered well situated within the larger society, because the parts of the smaller case unit were taken to be a representative part of the larger society.

What are the key components of the design?

Probably the most important single concept is that of the *case*. Stake (1995) says a case comprises people or programs, but not a problem, a theme, or a relationship because these are too abstract and lack boundaries. Yin (2003) describes five components he considers crucial for CSD: questions, propositions, analysis, linking of data to propositions, and criteria for interpreting the findings. Questions include the research question or questions, especially *how* and *why* questions. Propositions are the object that is to be studied in the case. They are, or are similar to, *purpose* or *hypothesis* or even *thesis*. A proposition explicitly states what will be studied and how it will be

judged to be successful or unsuccessful. In a single-case design, the unit of analysis can be another word for the case itself. The unit of analysis is what the research questions study. For example, for many language teachers, the unit of analysis is students. The fourth component of case design is linking data to the proposition. Unfortunately, the term *proposition* is not always made clear by Yin (2003). It could mean the same as research hypothesis in the sense that it means "the-thesis-you-are-trying-to-show-demonstrate-or-prove." Whatever *proposition* means, in requiring an explicit statement of how the data is to be linked to the outcome, Yin asks the TREE to state what would be a successful outcome (and, by implication, what would be an unsuccessful outcome). This linking would allow readers to decide if the case study were successful or not. The fifth component is the stating of criteria for success. Yin says this cannot be done with the precision of a cut score on a test, which can stipulate that "at or above 92 is pass while below 92 is fail." Although case study tends to be categorized as part of the qualitative research tradition, as with any design, case study design can accommodate either quantitative or qualitative data. Popular data collection instruments in CSD include: interviews, participant observation, tests, textbooks, field logs, as well as documents such as class syllabuses, handouts, and homework assignments. Consult Brown (1995) for an extensive list of types of data.

How this design might look visually

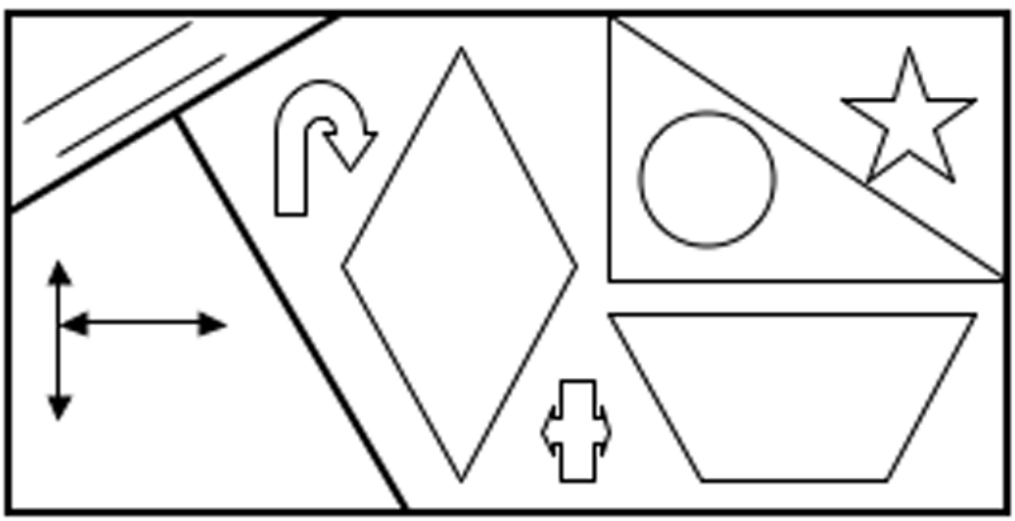


Figure 1. A visualization of a case study design showing a bounded case with three sections and multiple interior dynamics

Figure 1 shows one way to visualize a case study design. The heavy border indicates the case boundary, separating this case from other situations and other possible cases. In this image, the case is divided into three parts: one in the upper left hand corner, one in the lower left hand corner, and the rest. These three parts might represent various data collection modes, or they

might represent various dynamics in the case the study explains. Either way, this figure shows diverse multiplicity within the case. And yet for all the diversity, there is only one case, and it is this case that is under investigation.

According to Cohen, Manion, and Morrison (2000), the purpose of CSD observation "is to probe deeply and to analyze intensively the multifarious phenomena that constitute the life cycle of the unit with a view to establishing generalizations about the wider population to which that unit belongs" (p. 185). The terms *probe deeply* and *analyze intensively* may be taken to mean that CSD does not accept surface data at face value, but seeks through analysis to create a deeper explanation of the data. Multifarious phenomena may mean that whatever deeper explanation is created, it has multiple aspects in somewhat the same way a diamond has multiple reflecting surfaces. There is no way to know for sure if all of the phenomena have been investigated, but we can report the ones we found. Generalizations about a wider population to which the unit belongs may mean that the unit is of interest not only for its own sake, but for what it can tell us about some part of the world that interests us.

A brief scenario

Kiyoshi Yamada, whom everyone calls Ki (pronounced *key*), teaches Japanese in a U.S. university while working on his master's degree. In his first year class of 15 students, Ki has just administered his first major test, which in his opinion resulted in a rather strange distribution. Ten of the students scored in the middle (one standard deviation plus and minus), but three students scored very low (two standard deviations below the mean) and two students scored very high (two standard deviation above the mean). Why, Ki wonders, would that happen? Why do some students do so well and other do so poorly? Is it language aptitude, lack of studying, their over-extended social calendar, individual motivation, cultural affinity for the language, or some combination? One of the exceptional students was a Japanese American, but the other was not. Ki is considering a two-case design. One case would include the three students with the lowest scores, and the other case would be the two students with the highest scores. The general purpose would be to investigate why students fail and why students succeed. The proposition, to use Yin's (2003) term, would be that whatever was found in one group would be the opposite in the other. Since Ki has no specific theory in mind, and thus is not prepared to make any theoretical claims, this would be a descriptive study. The length of the study would be the current semester and perhaps the break following the semester if students were available and could be contacted.

Practical steps in getting started

Hamel (1993, p. 41) lists five (what he calls) practical processes in CSD, described in sociological terms. I recast these in research terms more familiar to second language teachers.

1. *Define the object of study.* Although for Hamel, CSD is part of a qualitative research tradition which allows the object of study to emerge gradually, he also favors a tradition that carefully defines the object of study before the study begins, or at least as part of the initiation of the study. In this sense, the object of study is the same as or similar to the construct.

2. *Select the case that will be the object of study.* Case selection is informed by an appropriate theory connecting the case to the construct.
3. *Decide on data collection methods*. This third step is fairly straightforward; Hamel reminds readers that data can include official documents, overheard remarks, and personal writing.
4. *Write*. This seems to include writing the analysis, as he discusses the difference between common language and the special language used in the report.
5. *Construct the explanation.* This appears to mean offering an explanation in terms of theoretical understanding.

Stake (1998), on the other hand, lists six steps:

1. Bounding the case and conceptualizing the object of study
2. Developing the research questions
3. Seeking patterns of data to develop the issue identified by the RQs
4. Triangulating the data for various interpretations
5. Selecting alternative interpretations to pursue
6. Developing generalizations about the case

What is the design good at?

There is an ongoing discussion between proponents of experimental design and case study design over the suitability of CSD for establishing cause-and-effect relationships. This discussion is philosophical in nature, with each side citing its sources (see Cohen, Manion, & Morrison, 2000, p. 181 for a citation explaining that CSD can establish a cause-and-effect relationship). Another source is Yin (2000) who argues convincingly that case study design can be used to specify and test rival theories because CSD is good at answering *how* and *why* questions. Less controversial is the claim that CSD is teacher friendly. McDonough and McDonough (199) cite Adelman (1991) as summarizing reasons CSD is user-friendly for teachers:

> Case Study Design is 'strong in reality,' allows for generalizations about an instance, or from an instance to a class, recognizes the complexity of 'social truths' and alternative interpretations, can form an archive of descriptive material available for reinterpretation by others, is a step toward action for staff or institutional development, and finally presents research in an accessible form. (p. 217)

Nunan (1992) observes that a major strength of case study design is its suitability for small-scale research of the type often done by teachers. One reason for this may be that individual students, groups of students, and classes are ready made, so to speak, for use as case studies. Finally, CSD can also be an option when a TREE has little control over events (Cohen, Manion, & Morrison, 2000, p. 182) or when context is important and events cannot be experimentally manipulated (Yin, 1993, p. 39).

What is it not so good at?

Users of CSD should be prepared to encounter some researchers' argument that CSD is "useful for purposes unrelated to inferring causation, such as assessing whether there was a treatment and how well it was delivered, or generating new hypotheses about the phenomenon under investigation" (Cook & Campbell, 1979, p. 98). The biggest drawback of CSD, according to Shadish, Cook and Campbell (2002) is the lack of counterfactual evidence from a control group. Counterfactual evidence is what might happen without experimental treatment. They continue to maintain that CSD should be used, but more for hypothesis generation than hypothesis testing (Shadish, Cook & Campbell, 2002, pp. 130, 501).

What issues and concerns does the design highlight?

Case study has had its ups and its downs. It arose in the early 20th century and was an important research model until the 1930s, when it lost ground to a new idea of theory and theorizing. This new idea was the development and appreciation of statistical methods combined with the desire in sociology to develop theory. Case study research and the field data it represented were rejected because they were not considered scientific and objective, and also because they involved researcher bias (Hamel, 1993). Specifically, it was argued that a single case could not be representative of the phenomenon under study, nor could the analysis be rigorous because of the researcher's subjective bias, in addition to the bias of the informant in the field on whom the researcher relied. The new theories of sociology required validation and generalizability, which statistics were seen as providing. The inductive, bottom-up theorizing of case studies was discredited in favor of a deductive, top-down confirmatory research approach that used statistics and an accompanying large N-size. As Hamel (1993) summarizes:

> The result was a reversal in the process by which field materials were constructed into sociological theory. In the Chicago School tradition, theory was built from field materials. In the statistical method, theory gave dimension to, and even validated, the representativeness of empirical data. (p. 21)

Starting in the 1940s, case study design entered an era of decreased popularity. Those in the new statistical wave considered it a secondary form of research, because its data was from the field, thus contaminated, biased, and not suitable for purposes of theory construction, theory validation, and generalizability. Case study design was considered exploratory; something similar to a pilot study, used for justifying a larger study with experimental design and statistics.

Cohen, Manion, and Morrison (2000, p. 183) maintain that although CSD frequently is associated with the interpretative tradition, it still must demonstrate reliability and validity. Is a single case generalizable? Atkinson and Delamont (1986) say no, while Simons (1980) says yes. McDonough and McDonough (1997, p. 216) cite Atkinson and Delamont as saying that CSD fails to build in features such as sampling or experimental treatment that would allow extrapolation to a wider population.

Further reading

In the second language acquisition literature, there are several well-known case studies: Hakuta (1976, 1986), a Japanese child learning English; Schumann (1978), a Spanish-speaking adult learning English; Wong Fillmore (1976), five children; Schmidt (1983), Wes, a Japanese in Hawaii learning English; and Schmidt and Fronta (1986); and Schmidt himself learning Portuguese for five months in Brazil. Johnson (1992, p. 77-82) lists, discusses, and gives examples of various uses of CSD investigating writing, SLA, and literacy. Other published articles described as case studies include Anderson (1998).

Some introductory research textbooks include chapters on case study design as well. These include Brown and Rogers (2002), Casanave (2010), Cohen, Manion, and Morrison (2000), Creswell (2003), Johnson (1992), McKay (2006), and Nunan (1992). For example, Brown and Rogers (2002) devote an entire chapter to an innovative case study design. They enable readers to experience case studies, first by engaging in a detailed linguistic analysis of the writing of Helen Keller, and second by engaging in a 24-hour case study conducted by the readers on themselves.

Some texts, however, are devoted entirely to case study design. These include Hamel (1993), Stake (1995), and Yin (1993; 2003). Hamel (1993) is short (76 pages), but helpful. It is a translation from French with a strong sociological orientation. This sociological tradition means that many concepts familiar in applied linguistics appear using new terminology. For instance, the more familiar word *construct* is now termed *the object of study* and *generalizability* is discussed as *representativeness*. Another problem is explaining new theoretical concepts that underlie case study using terms from mathematical theory. On the other hand, Hamel supplies an historical background that helps explain the political and methodological issues involved in how case study was discredited in the 1930 by proponents of survey design and statistical analysis.

The Making of a Teacher by Grossman (1990) uses CSD to investigate the impact of teacher preparation courses on teacher knowledge. Grossman interviews and observes six new K-12 (elementary through secondary levels) teachers, three who had taken education courses and three who had not. She documents her methodology and lists her interview questions for each interview in separate appendices.

McKay (2006) includes a relatively short, but interesting section on case study research. Its strength comes from having been written from the point of view of second language and classroom research by a knowledgeable and experienced researcher; the weakness, however, is that in discussing external validity, she misunderstands Yin (2003) on a crucial point, confusing experimental design with survey design. Yin (2003, p. 37) argues that both case study design *and* experimental design rely on *analytical generalization,* which generalize to theory, whereas survey design relies on *statistical generalization,* which generalizes to populations and universes. But in McKay (2006, p. 73), experimental design is substituted for survey design. This change can lead readers to the false conclusion that Yin (2003) is contrasting case design with experimental design instead of contrasting case design *as well as* experimental design with survey design. As a result, experimental design is erroneously described as generalizing by statistics to (by implication) populations rather than acknowledging that experimental design generalizes to

theory that can be used to generalize to other situations. McKay (2006) does not have a chapter on experimental design and does not discuss it extensively. She concludes her discussion of external validity by noting, correctly, Yin's position that case study involves analytic generalization that can be used to create theory.

Yin (2003) is the most complete text on case study currently in print, and any TREE interested in initiating a case study would be advised to read and study it in depth. At six chapters and 180 pages, it is rather short, but it is so full of advice, directions, and examples that it has the feel of a longer text. Chapters include: designing case studies, preparing for data collection, collecting evidence, analyzing the evidence, and writing the case study for publication.

DISCUSSION QUESTIONS

Write any questions you had while reading about case study design.

Reflection on Case Study Research Design

1. What is the attraction of case study design for classroom teachers in general and/or for you in particular?

2. What problems or issues would you anticipate in using case design?

Task. Find a published research paper that either explicitly states that the researcher used a case design, or in your opinion did so. Copy or download the article, answer the following questions, and bring it to class.

1. Note the title of the article, the author, the year, and the title of the journal.

2. How does the author communicate that the design is CSD?

3. Where did the research take place?

4. Can you describe the case?

5. Describe the question or questions the author was trying to answer.

6. Was the construct mentioned?

7. Were you convinced? Do you accept the conclusions of the author or authors?

8. Can this paper serve as a model for you to follow?

Glossary of Key Case Study Design Terms

Bias In CSD, the issue of bias concerns the subjectivity of the researcher and the informant. The criticism of the researcher can be due to lack of definition of the construct. For instance, the case investigates *motivation*, but the term is never defined in such a way that the reader can determine what is and what is not motivation. In fact, bias is a problem with all research designs, including experimental design where bias is discussed as rater bias or as a threat to validity.

Boundaries In general, boundaries are what separates one thing from another. Boundaries are used to identify a case from a larger context. An intact class may be designated a case because there is a clear difference (boundary) between it and other parts of a school program. Anything that has boundaries is easy to identify, and anything that does not have boundaries is harder to identify. For example, a class has boundaries (meeting time, classroom, student enrollment, teacher assigned) but a concept or theory such as communicative competence is more difficult to identify as a specific case because it is harder to identify its boundaries.

Case The unit of investigation for case study design research. The case is what the TREE researches. A case can be one object or one person, or it can be a group of persons. A case is anything that can be investigated that can be clearly defined and identified. What identifies or defines the case is called the *boundary*. Case study design can be one-case, two-case, or multiple case investigations.

Context Context refers to the case and everything involved in the case. The opposite of context is a *variable* used in experimental design to focus its investigation by stripping away what it considers extraneous factors. These extraneous factors are the context.

Document A document is a possible data instrument that is written and likely prepared by the institution where the case is located. Examples include email, syllabuses, reports, letters, calendars, schedules, textbooks, newspapers, and minutes of meetings. See Brown (1995) and Yin (2003) for discussion and examples of documents.

Longitudinal Longitudinal is related to the word *long*. A research investigation is longitudinal if data is collected over a period of time or gathered multiple times. The opposite of longitudinal is *cross sectional*. Cross sectional research means data is gathered one time only. Case study tends to be longitudinal.

Naturalistic Naturalistic refers to something occurring in an ordinary or usual way, and not studied in isolation. In this sense, naturalistic is the opposite of experimental. Naturalistic inquiry emphasizes gathering firsthand data that renders faithful and authentic accounts by researchers who were present (Schwandt, 2007). It is closely associated with participant-observation data collection.

Field Study On-site observation; field refers to the location of the research in the place where the activity is naturally occurring, for example a classroom, as opposed to a controlled

laboratory situation.

Participant Observer *Participant* refers to the meaning that people give to their own experience. *Observer* refers to systematic, close, and sustained examination and inspection of something.

Representative This term is used by Hamel (1993) to discuss CSD. Statistically-oriented researchers, for example, question the representativeness of CSD. They question whether one case can be indicative of the phenomenon under investigation. This issue is also discussed under the rubric of generalizability.

Triangulation Gathering two or more aspects of something in which one is interested. In most research designs, triangulation refers to multiple forms of data collection (more than one kind) or collection of data over time (on more than one occasion). Triangulation is attempted mainly for validating and strengthening the interpretation. The rationale is that multiple occasions of data are more likely to reveal underlying structures and involve less bias than single occasions. Triangulation applies to all sorts of designs, but is especially popular in case study design.

References for Case Study Design

Adelman, C. (1991). Talk on action research given at the Centre for Applied Research in Education (CARE), University of East Anglia (March).

Anderson, J. (1998). Managing and evaluating change: The case of teacher appraisal. In P. Rea-Dickins & K. Germaine (Eds.). *Managing evaluation and innovation in language teaching: Building bridges* (pp. 159-186). London: Longman.

Atkinson, P., & Delamont, S. (1986). Bread and dreams or bread and circuses? A critique of 'case study' in education. In M. Hammersley (Ed.). *Controversies in classroom research* (pp. 238-55). Milton Keynes, Philadelphia, PA: Open University Press.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thompson.

Brown, J. D. (1995). *The elements of language curriculum*. Boston, MA: Heinle & Heinle.

Brown, J. D. & Rogers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.

Casanave, C. P. (2010). Case studies. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in Applied Linguistics*. London: Continuum.

Cohen, L,, Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Merrill Prentice Hall.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. New York, NY: Teachers College Press.

Hakuta, K. (1976). A case study of a Japanese child learning English. *Language Learning, 26*, 321-351.

Hakuta, K. (1986). *Mirror of language: The debate on bilingualism*. New York, NY: Basic Books.

Hamel, J. (1993). *Case study methods*. Thousand Oaks, CA: Sage.

Hitchcock, G. & Hughes, D. (1995). *Research and the teacher: A qualitative introduction to school-based research* (2nd ed.). New York, NY: Routledge.

Johnson, D. M. (1992). *Approaches to research in second language learning*. New York, NY: Longman.

Kaufman, D. & Brownworth. (Eds.), (2006). *Professional development of international teaching assistants.* Case studies in TESOL practice series. Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

McDonough, J. & McDonough, S. (1997). *Research methods for English language teachers*. London: Arnold.

McKay, S. L. (2006). *Researching second language classrooms*. Mahwah, NJ: Lawrence Erlbaum.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Schmidt, R. W. (1983). Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. In N. Wolfson and E. Judd (Eds.). *Sociolinguistics and language acquisition* (pp. 137-174). Rowley, MA: Newbury House.

Schmidt, R. W. & Fronta, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.). *Talking to learn: Conversation in second language acquisition* (pp. 237-326). Rowley, MA: Newbury House.

Schumann, J. (1978). *The pidginization process: A model for second language acquisition*. Rowley, MA: Newbury House.

Schwandt, T. A. (2007). The Sage dictionary of qualitative inquiry (3rd ed.). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Simons, H. (Ed.) 1980. *Towards a science of the singular*. Occasional Papers 19, Centre for Applied Research in Education: University of East Anglia.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stake, R. E. (1998). Case Studies. In N. K. Denzin & Y. S. Lincoln (Eds.). *Strategies of qualitative inquiry* (pp. 86-109). Thousand Oaks, CA: Sage.

Tuckman, B. W. (1999). *Conducting educational research* (5th ed.). Orlando, FL: Harcourt Brace.

Wong Fillmore, L. (1976). *The second time around: Cognitive and social strategies in second language acquisition*. Unpublished doctoral dissertation. Stanford University.

Yin, R. K. (1993). *Applications of case study research*. Newbury Park, CA: Sage.

Yin, R. K. (2000). Case study evaluations: A decade of progress? In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.) (pp. 185-193). Boston, MA: Kluwer.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

CHAPTER SIX

ACTION RESEARCH DESIGN (ARD)

When teachers undertake research, it either changes teaching or it changes research. (Freeman, 1994)

In this chapter you will be exposed to the background of Action Research Design (ARD) and the considerations and steps involved in implementing an Action Research project.

Preview Questions

1. Do you see yourself primarily as a teacher or a researcher? Why?

2. Do you think teachers can also be researchers?

3. Are you more an active person or a reflective person?

4. Have you ever engaged in classroom research?

Introduction

Action research design is controversial, seemingly contradictory, and probably an as yet unfinished and still evolving design (Burns, 2005). At the heart of this approach is an attempt to join two apparent opposites: action and research. Classroom teachers tend to favor action over research; ARD is an attempt to put teachers in charge of their own research. The usual path for teachers to become involved in research is through a discipline other than teaching, such as applied linguistics, second language acquisition, or testing. There is nothing wrong with this, but the question remains: Who will research the areas and issues not covered by those disciplines but which are still of interest to teachers? ARD is one answer to that question.

Action Research Design defined

Field (1997) says, "The term 'action research' was adopted to describe a small-scale investigation undertaken by a class teacher" (p. 192). Patton (1990) defines action research as aiming "at solving specific problems within a program, organization, or community" (p. 157). ARD can be defined as small-scale investigation by teachers on specific classroom problems for the purpose of curriculum renewal and/or professional development (Field, 1997; LoCastro, 1994; Markee, 1996; Nunan, 1993; Patton, 1990).

Historical roots and key beliefs

According to Hitchcock and Hughes (1995), action research design in education began with Lawrence Stenhouse and the "Teacher as Researcher" movement in the 1970s. Kemmis and McTaggart (2005) give credit for the origin of action research to Kurt Lewin. Burns (1999) agrees

by saying that although the roots of ARD can be located in the work of John Dewey in the early 20th century, its origins are generally accredited to Lewin (1946).

As you can see from its historical roots, ARD is still fairly new and even now is being formed and discussed. We can think about ARD as lying on a continuum between curriculum renewal on one end and professional development on the other. Nunan (1993) points to teacher development as the purpose of action research. LoCastro (1994) suggests both curriculum renewal and on-going professional development. Field (1997) and Markee (1996) suggest course evaluation and innovation as purposes of ARD. For Wallace (1998), however, ARD is fundamentally a way of reflecting "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). Figure 1 captures the tension of these points of view.

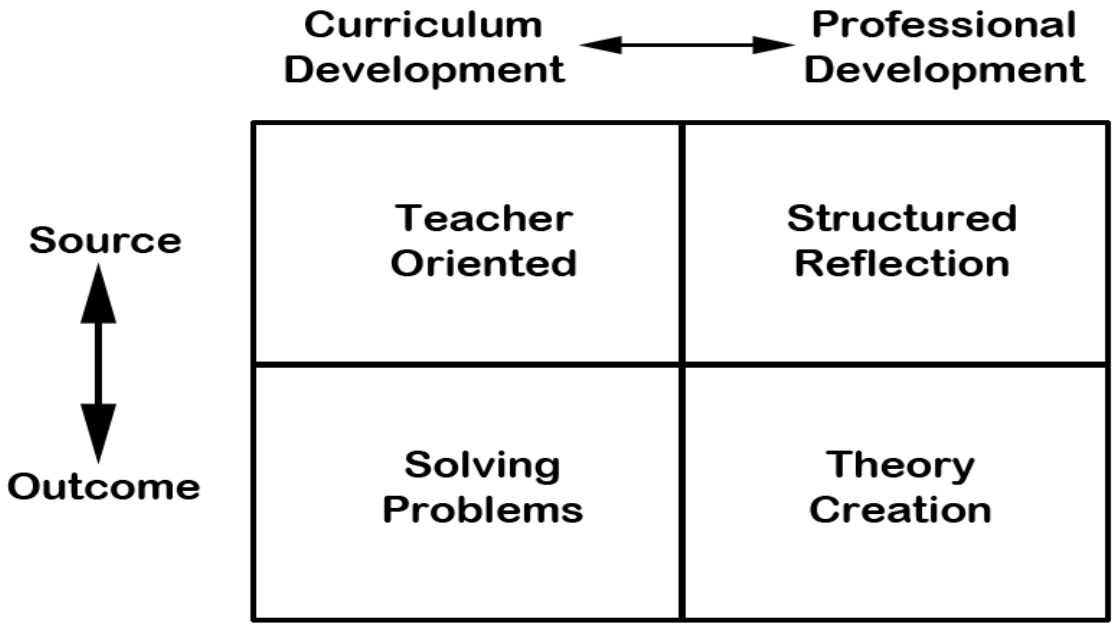


Figure 1. Focal points of action research showing the relationship between curriculum and professional development.

According to Figure 1, ARD can be understood as a source of curriculum renewal and also professional development. The action part is the curriculum renewal and the research part is the professional development. The source of curriculum renewal is the teacher, with the outcome being solutions to problems faced by the teacher. The source of professional development is structured reflection by the teacher, with the outcome being a new sense of freedom, emancipation, or empowerment resulting from the creation of teacher theory.

Teacher oriented means the teacher is central to curriculum renewal and evaluation (Nunan, 1992, p. 18; 1993, p. 41). According to this view, the classroom teacher's central position is the key distinction between ARD and other types of research. In ARD, the research questions come from the teacher's immediate concerns and problems (Crookes, 1991, p. 74). This means that ARD is usually small-scale investigation (Field, 1997) because the locus of the research is the teacher's intact class, and typically, intact language classes are not large.

Solving problems means that ARD is concerned with "solving specific problems within a program, organization, or community" (Patton, 1990, p. 157), which follows directly from the centrality of the role of the teacher. As Noffke (1997) shows, ARD has a long and complex history of differing strands and concerns, and this is where two of them emerge. The question of solving problems raises the issue of "whose problem?" Lewin (1946) represents the view of community action and social justice: The problems that need to be solved are those of minority groups with the help of the experts. For Brumfit and Mitchell (1990, p. 9), ARD is tied to the interests of teachers in the classroom. For them, the problems of the teacher need to be solved, not those of minority groups, such as students. This implies that in ARD, teachers identify a specific classroom problem to address (Hadley, 1997, p. 88) rather than problems derived from applied research theory. This might have implications for the literature review. If a teacher takes a problem from applied linguistics theory and application, the literature concerning that problem must be searched, synthesized, and discussed in the paper. If the teacher takes a problem from his or her own classroom experience, then the literature review may be reduced or even eliminated in favor of a description of the problem. Of course, the problem may already have been discussed by other teachers in published accounts.

Structured reflection is at the heart of ARD. Following Nunan (1993), Markee (1996), and Wallace (1998), reflection follows a general pattern of identifying a problem, gathering data, analyzing the data, forming or refining a hypothesis, creating a plan to test the hypothesis, implementing the plan, analyzing the data to decide what happened, looping if necessary, and finally making a report. Looping refers to repeating the process.

Theory Creation means that by using ARD, teachers empower themselves by constructing their own theory (Markee, 1996, 119; McNiff & Whitehead, 2006). Theory creation is a direct result of structured reflection because as a result of systematically collecting and analyzing data, reflection is promoted for teaching practice (Wallace, 1998, p. 4). Theory, as Widdowson (1993, p. 267) defines it, involves making ideas and beliefs explicit and systematic.

What are the key components of action research design?

Since ARD nearly always arises from a specific problem or issue in a teacher's professional practice, ARD involves the collection and analysis of data related to a problem in teaching for the purpose of discovery, reflection, and application to teaching (Wallace, 1998, p. 15). Nunan (1992, p. 18) identifies the key components of ARD as being initiated by a question, supported by data and interpretation, and carried out by a teacher in his/her own context. Markee (1996) gives six characteristics: 1) ARD is carried out by insiders; 2) uses any form of data (qualitative or quantitative); 3) is for the purpose of teacher behavioral and attitudinal change, 4) has no

expectation of generalizability; 5) seeks to improve classroom practice; and 6) aims at the development of teacher theory.

How action research design might look

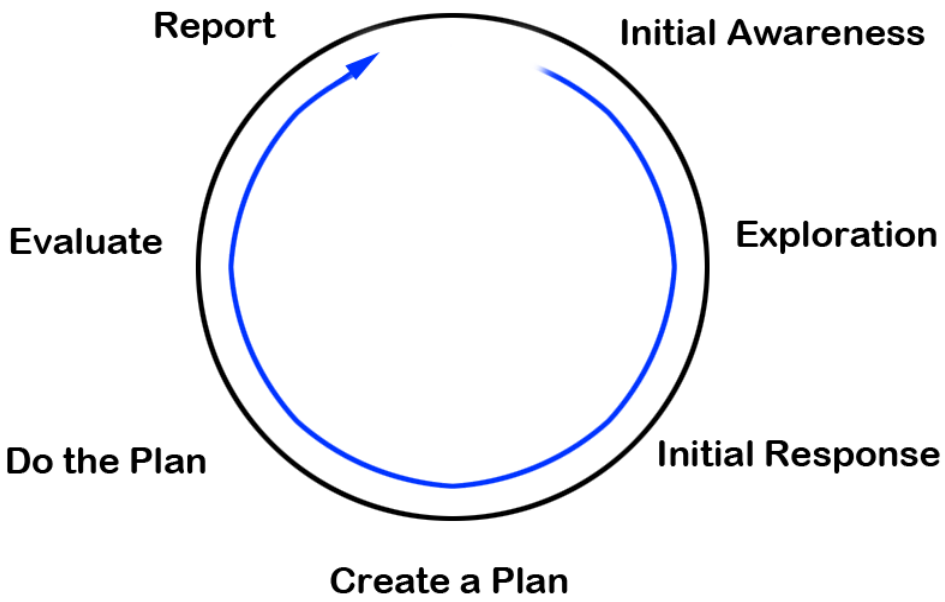


Figure 2. Action research design showing its research cycle

There is currently no consensus in action research design as to the number of steps, the order in which these steps should be taken, or even what the steps are. For example, Burns (1999) lists eleven steps, while Belleli (1993) lists only six. There is, however, a consensus that any and all steps of ARD should be taken as suggestive rather than prescriptive, reflexive rather than definitive, open-ended rather than fixed, and recursive. *Recursive* means that a step may be repeated and refined. Taken together, these characteristics indicate that it is up to each TREE to decide what to do and in which order to do it. This is a problem, on the one hand, because no particular step or series of steps can be recommended without knowing the context of the research under consideration. On the other hand, some guidance can be helpful, especially for TREEs new to action research design.

One solution is to group various steps into areas I will call *considerations*. A consideration is defined as a broad area consisting of many possible actions. A TREE can consult each area of consideration and decide which of the many possible steps should be done and in which order. In fact, a step might be appropriate in more than one area of consideration. For example, one recommended step is to consult with persons who might understand the problem and can offer

advice. It is hard to imagine any area of consideration in which this would not be a good idea. Seven areas of considerations are described: initial awareness, exploration, initial response, plan creation, plan execution, evaluation, and reporting the results.

Initial awareness The first general consideration in ARD can be called *initial awareness*, the TREE's response to noticing or becoming aware of a problem, puzzle, or a problematic situation. A *problem* is an issue a TREE becomes aware of, including classroom teaching. Characteristics of a problem are that it is noticeable, persistent, bothers the teacher (even if no one else is aware of it), and produces discomfort or even pain. In addition, a problem often affects multiple parties, including the teacher, but also others such as other teachers, administrators, and/or certain students. A *puzzle* is the same as a problem, but less intense. *Problematic* generally denotes that something is unsure, incomplete, or not well defined and connotes something that is wrong or unethical.

Exploration Four levels of exploration can be identified and discussed. The first level is general awareness; there is some awareness of a problem, but not necessarily more than that. What exactly the problem is, how complex it is, what its source is, what social and educational structures are involved, and what might be done to address it remain unclear. Level one of problem X may be typified by negative characterization, including *not enough* and *no control*. The TREE is aware of a problem situation, but complains there is not enough time, no money to deal with the situation, or that she does not have enough skill and/or experience. Additional problems might be that she has no control over the composition of the class, the administration, textbook selection, or some other key factor.

The second level of exploration is acknowledgement of external and internal sources of the problem. External sources of problems, puzzlements, and problematic areas might occur when someone complains to you about a problem he or she is having, and you recognize the problem as one you are also having; someone, perhaps a student, complains either directly or indirectly about you; an institutional review or evaluation identifies one or more problems, and you take the criticism seriously; someone mentions a problem that someone else is having, and it sounds familiar; or a group of teachers, including you, meets to discuss problems. Initially when you attended the meetings you were not sure these problems pertained to you, but as you listen and participate in the discussion, it becomes obvious that you have some or all of the problems. Perhaps you are taking a course, and in the midst of reading, listening, and participating in the course, a problem area is discussed which you recognize in your own teaching and learning. And finally, you attend a meeting or a conference presentation in which a problem is discussed, and as you listen, you realize this problem exists for you in your class. Internal sources of problems, puzzlements, and problematic areas might be: you begin to notice a situation in which people are suffering for reasons not of their own making; something in your class does not seem to be working well, or perhaps you are reading a book or article which points out a problem that you recognize as one you are having. At the second level, the TREE has two choices. If she ignores the external or internal prompts, she remains at level one. If she acknowledges the prompts, she has a clearer awareness of the problem.

The third level of exploration is decision. The TREE decides to engage in some form of systematic

investigation and research including but not limited to an action research project. This decision has the effect of giving the TREE permission to deal with a limited number of the manifestations of the problem. The TREE can accept, select, or formulate one problem among many to focus on.

The fourth level of exploration is the ethical dimension. At this level, the problem involves values and their violation. If there is no violation of an educational value, there is no problem. Imagine two teachers, both of whom have students who on a regular basis enter their class several minutes late. One teacher notices this behavior, but ignores it, saying that he understands that occasionally students are late, and that it can't be helped. For this teacher, no educational value is being violated. The second teacher, on the other hand, is irritated and over time, becomes increasingly distressed. This shows that a situation is not a problem until an educational or ethical value is involved and violated in some way. Signs that a value has been impinged on include irritation and aroused emotions; for example, you are angry or upset. Identification of a value may make a problem clearer and thus a solution clearer. It also means that action research is always closely aligned with a teacher's values and beliefs.

Steps in exploration will be separated by bullets rather than numbered to avoid implying a priority or sequence. Typical exploration steps mentioned in the ARD literature (in no particular order) include:

- Identify a problem or problems. Jot down as clearly as possible what you think the problem is, keeping in mind that your thinking will probably change as the research continues. Number and date your problem identifications so you can trace your understanding over time.

- Evaluate the problem. How serious do you consider the problem to be? What seems to be involved? Who are the affected parties?

- Experiment by trying something different. Ask yourself, "What happens when I ___?" (Curry, 2006).

- Engage in class observation, for example, by keeping a diary to document more exactly what is going on and/or what the problem is.

- Write a quick memo explaining the problem's history from your point of view. Write where you think the problem came from and how you first noticed it.

- Compose a more complete second memo with your analysis of the problem by following the seven levels of problem analysis described previously.

Initial response The third general area of consideration in ARD is the *initial response*. Initial response is any response a TREE undertakes which addresses the exploration (described in the exploration consideration), and which is usually above and beyond normal teaching activity.

Typical initial response steps include:

- Decide what action to take.

- Brainstorm all possible solutions to the identified problem. The assumption is that we always have more than one possible solution, although some are more suitable than others.

- Locate other teachers and persons to help you understand your problem and how to respond; ask them for their insights.

- If you haven't already done this already, create an initial hypothesis. A hypothesis is a guess or statement about what the TREE believes the problem to be, what causes it, and what may be done about it. An initial hypothesis is the first guess, or speculation, regarding what the problem is. This is done prior to actual investigation, which might lead to a revision of the hypothesis.

- Begin to gather what Nunan (1993) calls baseline data. The purpose of baseline data is to understand the situation before changes are implemented.

- Initiate a literature review. Does this problem have an agreed upon name that you can use to search for published articles and books? Have others dealt with it? If so, what did they find?

Create a plan The fourth general consideration in ARD is *create a plan*. Creating a plan entails generating an outline of possible actions to take. Earlier we saw that any set of ARD steps should be considered reflexive, open ended, and suggestive rather than prescriptive. Plan creation is a good example of these suggestions because creating a plan covers initial ideas of what to do, as well as the revising of those ideas based on new information. The plan is always open to change and modification as the research proceeds. Here are some typical steps to create a plan-- don't try to do them all. Select some that you think are do-able and appropriate for you at this stage of your action research.

- If you haven't already, create an initial hypothesis of what you think is happening.

- If you haven't already, list possible research questions. If this is not yet possible, list areas of interest that merit looking into. This might turn into researchable questions. If you have already listed possible research questions as part of your initial response, you may have changed your thinking. Review your first RQs by asking yourself if they still reflect what you want to investigate.

- Continue to deepen your literature review.

- Give some thought to the kind of data you might collect and how to collect it.

- For each possible solution, describe what blocks it. A block is anything that inhibits further action. An example of a block is deciding to do something, and then realizing that you have little or no idea of how to begin. Knowing that "how to begin" constitutes a block means that you can search for a first step. If you can't identify a block, focus on the solution. If you can identify a block, focus on how to deal with the block.

- Write yourself a memo after each consideration. Writing a memo not only reminds you about what you know, but it also starts the writing process, which is too often left until the end of the research.

Do the plan and evaluate the plan The fourth general consideration in ARD is *do the plan*. Implementing an action research plan—or a research plan involving any kind of research design for that matter—sounds easier than it is. Doing the plan assumes that a plan exists and that it is sufficiently detailed so that it can be done. Typical steps in doing the plan include:

- Create a plan for your research.
- If you have a plan, check it for a clearly stated purpose, actionable research questions, and ideas for data collection and analysis.
- Include a timeline of what you will do and when you will do it.
- Create an evaluation plan that states how you intend to evaluate your action plan.
- Look again at the research questions. Are they specific enough? Can they be answered? Do you know how to answer them? Are there too many? Not enough?
- Think ahead to reporting, and estimate how you will report the results of your research.

Report results The last general consideration in ARD is *reporting results*. This entails finding ways to present your findings both orally and in written form. Typical reporting steps include:

- Present your findings to colleagues. Seek a friendly but professional forum. This might be a teacher meeting in your school or a teacher meeting in your city, state, or province. If such a group exists, they may be looking for presenters; make yourself available by volunteering to present your results.
- Submit a proposal at a conference. If it is your first conference presentation, consider a local meeting as a rehearsal; think later about presenting at a national conference.
- Consider writing up the results of your research.
- Decide on the next round of investigation.

What are the advantages of action research design? Generally, ARD can be thought of as structured reflection that is natural and doable. It takes as its starting point teacher interest, which the teacher is most likely aware of. ARD seems to focus on what works best with particular students in particular settings (LoCastro, 1994, p. 5). More specifically, ARD has three important benefits:

1. It encourages teachers to reflect on their practice. therefore, it could lead to change.
2. It empowers teachers by releasing them from ideas handed down by past experience (Field, 1997, p. 192).
3. It allows teachers to receive ideas from academic researchers and trainers, and to implement those ideas. This helps make individual teachers authorities on the ideas and their implementation, rather than just passive receivers of ideas or practices.

What are the disadvantages of action research design? Nunan (1990) maintains that ARD "lacks the rigor of true scientific research" (p. 64). Generally, he claims, findings based on ARD have to be very careful about generalizability because other uncontrolled factors may be involved (1992, p. 19). A second problem teachers using ARD face is stating their problems as researchable questions. It is one thing for teachers to become aware that there are problems, but it is another thing to be able to formulate a problem as a researchable question. A researchable question is one that can be answered. Any teacher following any research design confronts this problem, sooner or later, but it seems that ARD is particularly vulnerable because it is advertised as being teacher friendly. This stance may lull teachers into thinking they do not have to articulate answerable research questions. In fact, it might be that the problems facing teachers using ARD are the same as those faced by teachers using any other research design, including problems with data collection, data analysis, literature search, instrument creation, and instrument validation. A third problem area is that there is little or no agreement on how to report findings. The findings of research from ARD may not be disseminated beyond a local program or organization. In many instances, there may not even be the expectation of a full written research report. Rather there may be briefings, staff discussions, and oral communication (Patton, 1990, p. 157). This wide range of reporting formats can be divided into four areas: verbal reports, presentations, reports, and articles each of which can be seen to have an informal and formal aspect. These formats can be seen summarized in Figure 3.

Format	Informal	Formal
Verbal report	Casual conversation with fellow teachers and administrators.	Short report at a staff or teacher's meeting.
Presentation	A report at a teacher session at a conference.	A conference meeting; a poster with handouts.
Report	Notes handed out at a teacher meeting.	A newsletter article.
Article	A pedagogical article or a short report of research.	An academic journal article.

Figure 3. Action research reporting format categories showing informal and formal examples

Informally, verbal reports may be conversations with other teachers and administrators in chance meetings or over coffee. More formally, they may consist of short reports at a teacher

meeting. Informally, presentations may be a report at a teacher meeting or a poster session at a conference, and more formally may be a peer reviewed presentation at a conference. Informally, memos may be brief notes handed out at a teacher meeting, and more formally memos may be several pages of prose that could include tables and references. What I am calling a report could even be a short newsletter or newspaper article. Informally, articles may be pedagogical, or a research report, and more formally may be a full-length research article in a journal. All these are considered legitimate forms of reporting AR research.

What are the claims of ARD?

There are two major roles for ARD: exploratory and emancipatory (Ellis, 1997, p. 86). Ellis also claims that ARD can be used either for curriculum development or teacher awareness (1197, p. 87). McNiff and Whitehead (2006) claim that ARD allows teachers to think and act for themselves. Writing from a sociological perspective, Lewin (1947) describes field experiments—which he accepts as the norm for research—and claims the most difficult task of research is to maintain a balance between the objective and subjective aspects of data interpretation. By objective he means the close description of what occurred so that pretest and posttest measurements can be taken. By subjective, he means "the initiative shown by individuals and subgroups" (Lewin, 1947, p. 151); in other words, the values and goals of not only the participants, but also the researchers themselves.

Many different data collection instruments can be used--for example, progress tests, field notes, diaries, questionnaires, interviews, introspection, peer observation, and recording lessons (Field, 1997, p. 192). Creswell (2002) suggests brainstorming to decide which types of data can be collected and which would shed light on the research questions.

Issues and concerns One issue that concerns action research practioners is the dissemination of action research results. How are the results of problem solving or theory creation to be actually acted on, and not be published or presented and then forgotten? In other words, how is the action in action research to be accomplished? A second issue is whether ARD is primarily research or primarily reflection. Wallace (1998, p. 17) says that ARD is a form of structured reflection, and that while it is true that for reflection to have value it must be valid, *the process* of the research is more important than *the product*. Process seems to emphasize development, either personal or professional, while product seems to emphasize knowledge and its use.

A third issue is, to the extent that ARD is research, how good is it? In one sense, this is a silly question because there is a lot of poor research produced through other research designs. But in another sense, it is a relevant question because of the way some researchers describe ARD. For example, Patton (1990, p. 157) says that because ARD is specific to a problem, the research methods may be less systematic and more informal. "Less systematic" and "more informal" might be taken as code words for less than rigorous research. Other researchers stress that the results of ARD are not generalizable beyond the situation in which they were generated (Nunan, 1992, p. 19). Ellis, (1997) says when ARD is local and no generalization is attempted, there is no need to conform to the objectives and methods of proper research. Crooks (1993) argues that within ARD, it is less necessary to attend to reliability and validity or to academic style. This view

can be contrasted with that of Brumfit and Mitchell (1990) who point out, "[T]here is no good argument for action research producing less care and rigor unless it is less concerned with clear understanding, which it is not" (p. 9). However, Ellis (1997, p. 87) concludes that if teachers use ARD to develop awareness, then process is more important. On the other hand, if teachers use ARD for curriculum development, then product is important and reliability and validity need to be established, although there is no need to attend to generalizability of findings.

Further Reading

Only a few research methods textbooks include chapters on ARD. Those that do include *Research Methods in Education* (5th ed.), by Cohen, Manion, and Morrison (2000); *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research,* by Creswell (2002); and *Educational Research:* and *Competencies for Analysis And Application* (6th ed.), by Gay and Airasian (2000).

Some texts, however, are devoted entirely to ARD. *Collaborative Action Research for English Language Teachers,* by Anne Burns (1999), is 254 pages long and emphasizes how teachers can work together. It is also rich in reports of teacher experience. *Introduction to Action Research: Social Research for Social Change,* by Greenwood and Levin (2007) has 300 pages. The authors are an American scholar and a Norwegian engineer; the book covers ARD from multiple points of view, including academic and industrial. It includes chapters on the history of ARD as well as the epistemological basis of ARD. *All You Need to Know about Action Research* is by two researchers from the United Kingdom, Jean McNiff and Jack Whitehead (2006). This book has an extensive reference section and a glossary. Despite its practical-sounding title, this book is philosophical and theoretical in a helpful way; there are chapters on the assumptions of action research (chapter three) and the history of action research (chapter four). *Action Research: Teachers as Researchers in the Classroom,* by Craig Mertler (2006), is directed at K-12 teachers. With nine chapters and 251 pages, Mertler discusses many steps at the basic level, defines many terms, and lists web sites. His literature base, that is, the number of works he consulted, however, is small. *Action Research* (3rd ed.), by Australian researcher Ernest Stringer (2007), has 277 pages and nine chapters, including ones on getting started and the role of theory in ARD. The approach of this text defines the role of the researcher not as an expert who does research, but rather as a facilitator who helps others define their problem and work toward solutions. *Action Research for Language Teachers,* by Michael Wallace (1998), has ten chapters and 273 pages. This text is written for language teachers and has helpful chapters on thinking about ARD (chapter one) and how to get started (chapter two).

DISCUSSION QUESTIONS

Write any questions you had about ARD.

Reflection on Action Research Design

1. What is the attraction of action research design for classroom teachers in general and/or you in particular?

_____

_____

2. What problems or issues would you anticipate in using this design?

_____

_____

Task 1. Find a published research article that either explicitly states that the researcher used an action research design, or an article in which the researcher did not explicitly state this, but you believe used an action research design. Copy or download the article, answer the following questions, and bring the article and your answers to class.

1. State the title of the article, author, year and the title of the journal.
2. How does the author tell you that the research is ARD?
3. Does the author(s) state the purpose of the research? Are there research questions?

4. In your opinion, is this actually an example of ARD? What actions taken in the article communicate to you that ARD was the model?

5. Were you convinced? Do you accept the conclusions of the author or authors?

6. Can all or part of this paper serve as a model for your research?

Task 2. To get started, consider a class, real or imaginary, you have taught, are teaching, or might teach. What is a problem or problems that you are aware of?

_____

_____

_____

_____

Task 3. Define the problem mentioned in task 2 in terms of situation, ethical norm, possible solution, and a block to that solution.

_____

_____

_____

_____

Glossary of Key Action Research Design Terms

Awareness Composed of noticing, investigating, and tentative identification of something. Noticing usually concerns a problem or previously unrecognized data. Investigating involves an initial scope, for example, "Is it just me, does it concern all my students, or even the whole school?" Tentative identification includes early identification and/or action. Thus, to become aware of a phenomenon in a classroom means to notice it, notice whom it effects, and, if possible, to give it a name.

Change process The way teachers implement innovation. For some (Markee, 1997) change and innovation are the same, but for others innovation is planned and managed, while change is what happens over time.

Classroom research Any type of research using any design, but which is centered on the classroom and the teacher in that classroom. Classroom research is assumed to be initiated by, or strongly controlled by, the classroom teacher.

Critical The belief that all forms of educational practice, including research practices, have hidden political assumptions and benefits which serve some group's interest. In other words, nothing is neutral.

Curriculum renewal An ongoing process of evaluating a curriculum that includes systematic evaluation or innovation.

Emancipation An issue on the research or professional development side of the focal points of action research (see Figure 1). The goal of emancipation is to be free. For a person, say a teacher, to be free in the context of an institution and a classroom with particular students, requires awareness of his/her situation (structured reflection) combined with a self-ideology (theory creation) plus action to implement the ideology.

Empowerment According to Kemmis and McTaggart (2005), empowerment can be understood as "a capacity for individuals, groups, and states to interact more coherently with one another in the ceaseless process of social reproduction and transformation" (p. 594). Empowerment is not political activism or getting power; rather, it is the result of communication and interaction that results in decisions. Empowerment means having a say and being able to participate in the decision making process. The opposite of empowerment is silence, nonparticipation, and victimization.

Informal methods Hubbard and Power (2003) describe informal interview methods, which can be taken for informal methods in general, in the following way: They often arise spontaneously in the midst of regular classroom activity and are part of the usual question and answer of teacher-student interaction. In other words, informal methods refer to those ways of gathering data that would have or might have happened normally, or at least would appear that way to a visitor to the class.

Innovation Markee (1997) offers language teachers a full discussion of curricular innovation, and suggests that innovation is a managed or conscious process of change involving materials, skills, and values. Innovation can be good or bad because the results of any innovation are not known until later in the process.

Participatory A stance toward research that opposes a view or understanding of the researcher as unbiased, objective, and separate from those being researched. As Bishop (2005) says, for researchers following a participatory understanding:

> [T]his approach means that they are not information gatherers, data processors, and sense-makers of other people's lives; rather they are expected to be able to communicate with individuals and groups, to participate in appropriate cultural processes and practices, and to interact in a dialogic manner with the research participants. (p. 120)

Reflection Richards and Schmidt (2002) define reflection as having three parts: 1) thinking back, remembering, or recalling; 2) considering the experience; and 3) trying to better understand. They point out that reflection is considered important for learning, and that reflective exercises are now common in teacher education as well as regular classroom activities. Mertler (2006) defines reflection as critically exploring what you did, why you did it, and what effect it had. It is worth noting that language research using any type of research design induces reflection.

References for Action Research Design

Belleli, L. (1993). How we teach and why: The implementation of an action research model for in-service training. In J. Edge & K. Richards (Eds.). *Teachers develop teacher development: Papers on classroom research and teacher development* (pp. 65-75). Oxford: Heinemann.

Bishop, R. (2005). Freeing ourselves from neocolonial domination in research: A Kaupapa Maori approach to creating knowledge. In N. Denzin & Y. Lincoln (Eds.). *The Sage handbook of qualitative research* (3rd ed.) (pp. 109-138). Thousand Oaks, CA: Sage.

Brumfit, C., & Mitchell, R. (1990). The language classroom as a focus for research. In C. Brumfit & R. Mitchell (Eds.). *Research in the language classroom* (pp. 3-15). London: Modern English Publications in association with the British Council.

Burns, A. (1999). *Collaborative action research for English language teachers.* Cambridge: Cambridge University Press.

Burns, A. (2005). Action research. In E. Hinkel (Ed.). *Handbook of research in second language teaching and learning* (pp. 241-256). Mahwah, NJ: Lawrence Erlbaum.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

Greenwood, D. J., & Levin, M. (2007). I*ntroduction to action research: Social research for social change* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, NJ: Merrill Prentice Hall.

Crookes, G. (1991). Action research for second language teachers—It's not just teacher research. U*niversity of Hawai'i Working Papers in ESL 10(2),* 73-90.

Crookes, G. (1997). SLA and language pedagogy: A socioeducational perspective. *Studies in Second Language Acquisition, 19*(1), 93-116.

Curry, M. J. (2006). Actions research for preparing reflective language teachers. *TESOL HEIS News* (Newsletter of Higher Education Interest Section) *25(1),* 1-4.

Ellis, R. (1997). SLA and language pedagogy: An educational perspective. *Studies in Second Language Acquisition, 19*(1), 69-92.

Field, J. (1997). Key concepts in ELT: Classroom research. *English Language Teaching Journal, 51(2),* 192-193.

Freeman, D. (1994). Doing, knowing, and telling: Research and what teachers know. *The Language Teacher, 18(9),* 6-7.

Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Upper Saddle River, NJ: Merrill, Prentice Hall.

Hadley, G. (1997). Action research: Something for everyone. In D. T. Griffee & D. Nunan (Eds.). *Classroom teachers and classroom research* (pp. 87-98). Tokyo: The Japan Association for Language Teaching.

Hitchcock, G., & Hughes, D. (1995). R*esearch and the teacher: A qualitative introduction to school-based research* (2nd ed.). New York, NY: Routledge.

Hubbard, R. S., & Power, B. M. (2003). *The art of classroom inquiry: A handbook for teacher-researchers* (2nd ed). Portsmouth, NH: Heinemann.

Kemmis, S., & McTaggart, R. (2005). Participatory action research: Communicative action and the public sphere. In N. K. Denzin & Y. S. Lincoln (Eds.). *The Sage Handbook of qualitative research* (3rd ed) (pp. 559-603). Thousand Oaks, CA: Sage.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, 2, 34-46.

Lewin, K. (1947). Frontiers in group dynamics: II. Channels of group life; social planning and action research. *Human Relations I,* 143-153.

LoCastro, V. (1994). Teachers helping themselves: Classroom research and action research. *The Language Teacher, 18(2),* 4-7.

Markee, N. (1996). Making second language classroom research work. In J. Schachter & S. Gass (Eds.). *Second language classroom research: Issues and opportunities* (pp. 117-155). Mahwah, NJ: Lawrence Erlbaum.

Markee, N. (1997). *Managing curricular innovation.* Cambridge: Cambridge University Press.

McNiff, J., & Whitehead, J. (2006). *All you need to know about action research*. Thousand Oaks, CA: Sage.

Mertler, C.A. (2006). *Action research: Teachers as researchers in the classroom.* Thousand Oaks, CA: Sage.

Noffke, S. E. (1997). Professional, personal, and political dimensions of action research. In M. W. Apple (Ed.). *Review of research in education 22* (pp. 305-343). Washington, D.C.: American Educational Research Association.

Nunan, D. (1990). Action research in the language classroom. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 62-81). Cambridge: Cambridge University Press.

Nunan, D. (1992). *Research methods in language learning.* Cambridge: Cambridge University Press.

Nunan, D. (1993). Action research in language education. In J. Edge & K. Richards (Eds.). *Teachers develop teachers research: Papers on classroom research and teacher development* (pp. 39-50). Oxford: Heinemann.

Nunan, D. (2005). Classroom research. In E. Hinkel (Ed.). *Handbook of research in second language teaching and learning* (pp. 225-240). Mahwah, NJ: Lawrence Erlbaum.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Pugh, K. J., & Bergin, D. A. (2005). The effect of schooling on students' out-of-school experience. *Educational Researcher, 34(9),* 15-23.

Richards, J. C., & Schmidt, R. (2002). *Dictionary of language teaching & applied linguistics* (3rd ed.). Burnt Mill, Harlow: Longman.

Stringer, E. T. (2007). *Action research* (3rd ed.). Thousand Oaks, CA: Sage.

Wallace, M. J. (1998). *Action research for language teachers.* Cambridge: Cambridge University Press.

Widdowson, H. G. (1993). Innovation in teacher development. *Annual Review of Applied Linguistics*, *13,* 260-275.

PART THREE
Data

INTRODUCTION TO DATA COLLECTION INSTRUMENTS (DCIs)

The importance of data

Data is the lifeblood of research. Data connects theory (ideas about the world) to practice (the world). Without data, we have to take the researcher's word for whatever claims she is making. Data allows us to look over the researcher's shoulder and see what he saw. Data makes research *empirical*, and empirical research is highly valued because it represents something outside our opinions and ourselves.

What is a Data Collection Instrument?

Although *data* is the plural form and *datum* is the singular, I follow the practice of Brown (2001) by using *data* for both singular and plural meanings. Examples of data collection instruments include questionnaires, various types of tests, observation schemes, and transcription protocols for speech samples. In general, a data collection instrument (DCI) can be defined as the means, either physical or nonphysical, by which data is produced. More specifically, a DCI can be defined as the means (physical or nonphysical) of producing quantitative or qualitative data to be analyzed and interpreted. This process requires some kind of validation evidence that the data is sufficiently related to the construct of interest. Figure 1 illustrates this definition, beginning with the right arrow pointing at *produces data*. The purpose of any DCI is to produce data measuring a construct. The data must be analyzed and interpreted, and the resulting interpretation then requires validation evidence.

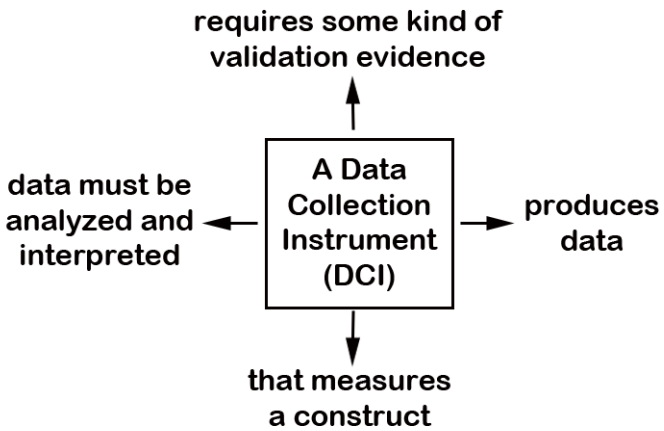


Figure 1. The purpose and function of a data collection instrument

Reflecting on the definition of a DCI, *physical* could take the form of printed words (a test, a questionnaire); *nonphysical* could take the form of a procedure in a TREE's mind (such as an

interview protocol). Data can be quantitative, referring to numbers (for example, test scores) or qualitative, using words (for example, observation notes). Any DCI can produce quantitative or qualitative data. For example, a questionnaire could produce numbers through closed-ended items such as ratings scales, or words by the use of open-ended items, such as questions of opinion. Raw data, in and of itself, is useless; it must be analyzed and interpreted to be valuable for research purposes. *Analysis* refers to the process by which a large amount of raw data is reduced; *interpreted* refers to the assigning of meaning to the reduced data. *Validation* is an argument by which evidence shows that the analyzed interpretation based on data to some extent reflects the construct.

Examples of data collection instruments

Here are some popular DCIs, along with a reference to a source that discusses it, and a short description.

- *Conversation transcriptions* (Tuckman, 1999, p. 412) are the written form of a spoken conversation. The conversation is written in order to show the aspects of what was said, so that it can be analyzed and studied. Various linguistic analytical categories are used, such as length of pauses, number of words, types of words, and backchanneling.

- *Criterion-referenced tests* (Brown & Hudson, 2002; West, 1994, p. 8) are typically teacher-made tests that measure mastery of specific material. Examples of criterion-referenced tests are weekly quizzes, midterms, and final examinations.

- *Delphi technique* (Brown, 1995, p. 49) is a protocol used when expert participants can't meet face to face, but want to reach a consensus. Questions are sent to the first person with a request for an answer and an explanation. The first person's comments are sent to the second person with a request not only for an answer, but also comments on the first person's comments. This is called a round and continues until all persons on the list have seen all the other persons' comments. Then round two begins and continues until a concensus has been reached. A Delphi may use only written input, such as a questionnaire or a combination of written and interview data.

- *Diaries or journals* (McDonough & McDonough, 1997; West, 1994, p. 8), also known as logs or letters, are usually written by participants in a study to record data, thoughts, ideas, and feelings about teaching or learning.

- *Documents* (Yin, 2003) are written instruments, sometimes prepared by a school or organization. These could include syllabi, schedules, minutes of meetings, letters, email, or evaluation reports. Sometimes documents are prepared by outside agencies, for example newspapers, magazine articles, pamphlets that contain institutional purpose, goals, and direction.

- *Interviews* (Griffee, 2005; Kvale, 1996; Tuckman, 1999, p. 403; West, 1994, p. 7) are usually one-to-one face-to-face meetings in which the data-gatherer asks questions to someone being interviewed.

- *Observation* (West (1994, p. 7) is the act of watching something and recording the results in a way that produces data that can be analyzed and interpreted.
- *Performance tests* (McNamara, 1997) are tests that judge actual experiences of persons doing tasks that they are likely to do in real life.
- *Placement tests* (Brown, 1996) can be paper- or computer-based tests that assign the test-taker a proficiency score. The test results, perhaps with other data, is used to assign the candidate to a program level matching their proficiency.
- *Questionnaires* (Brown, 1995, p. 50; West, 1994, p. 7) are often paper- or computer-based instruments asking respondents for their opinions, as opposed to measuring learning.
- *Text analysis* (Connor, 1994; Swales, 1994, 2004) is the analysis of certain genres or types of text to enable students to understand how the text works. For example, Swales analyzed the genre of research papers, especially introductions to research papers, and found that there was a three-step process.

A data collection instrument measures a construct

A DCI produces data that is the manifestation of a construct. A construct, according to Richards, Platt, and Platt, (1992, p. 80) is a concept that is inferred, based on observable phenomena. A construct can help in the analysis and understanding of events and phenomena. In other words, the construct is what the TREE tries to measure, and the data collection instrument is the means of doing so. Although a DCI collects the observable phenomena or data, a DCI is attempting to measure a construct. Therefore, a TREE must identify and define the construct by means of theory if available, or his/her own definition, if a theory is not available.

A data collection instrument produces data

Data in various sciences can come from almost anything--including plants, pictures from a telescope, or recorded sounds. In language research, for example applied linguistics, educational linguistics, or second language acquisition, the data we gather tends to be numbers or words. As mentioned previously, numbers are quantitative data and words are qualitative data.

Test scores are an example of quantitative data. These scores, called raw data, can be reduced to categories, such as averages or means, and interpreted as an indicator of proficiency (the construct) or as the level of mastery of certain information (the domain).

An example of qualitative data collection might be interviews from members of language class. The data might take the form of notes or a very complete transcription of the interview (again, raw data), which must be analyzed.

Data must be analyzed and interpreted

Data does not speak for itself. Picture walking down the hall of your school and meeting a colleague. You say “hello” and he starts reciting lists of numbers. After a few minutes, your colleague stops

saying numbers and continues on his (or her) way. What would you make of this? Possibly your colleague is delusional, suffering from a brain abnormality in which numbers are substituted for words, or perhaps your colleague is playing a joke on you. Either way, the numbers by themselves do not make sense. That is because data does not make sense by itself; data only makes sense in the context of a research design, theory, purpose, and research questions. We need to know what the data means, and the data by itself can't tell us, only a meaningful context in which we can interpret the data can tell us.

A data collection instrument must be validated

It is important that we do not automatically trust data. We should ask where it came from, how it was collected, and who collected it. We should also ask to see the collection instrument or instruments. There should also be questions about the data's reliability, how this reliability was determined, and especially how the conclusions were validated. We should not accept the results of one study as final no matter how convincing the data may be (Campbell, 1969). If the results of the study are replicated over a period of time, and many studies show the same result, perhaps we can believe that the conclusions have merit. In other words, we need to be skeptical about data and data collection instruments. All instruments must be validated. (Actually, this is not correct. It is not the instrument that must be validated; rather, it is the *interpretation* that must be validated.) Although validation is considered a unitary process, it is often discussed under two rubrics: reliability and validity. The process of demonstrating that the raw data is consistent from one data collection episode to another is called *reliability*. The argument that there is a fairly strong connection between the interpretation and the construct is called *validation*.

Research designs and data collection instruments

Using the human body as a metaphor, design is to the skeleton as data is to the flesh. To refer to design as the skeleton of any research paper refers to the ability of the design to shape the research process. *Shape* means how various parts of the research and the resulting research article are arranged, and how they relate to each other. For example, if a TREE uses a variation of experimental design, then we can expect to see the objects of interest identified as *variables*, and we can expect to see those variables measured in some way. To refer to data as the flesh of a research paper refers to the ability of data to determine the appearance of the research report. *Appearance* refers to the way in which data is collected, analyzed, and interpreted. For example, in some designs, statistics are used because that design is closely associated with sets of quantitative data as compared to each other. Thus, just as the flesh interacts with the skeleton, data interacts with the design to form typical or characteristic contours. This is why we often use the phrase *skin and bones*; they go together. It is hard to imagine a person with skin but no bones, or bones with no skin. Considering this metaphor, we can see why research papers that use the same design tend to resemble each other.

Designs themselves are not usually seen and seldom cited, but data collection instruments and the data they produce are very visible. When we read a research article in a journal, the data captures our attention. One result of this imbalance of awareness between design and data is that design tends to be ignored and data emphasized. This may be why data collection instruments

are sometimes substituted for design. By *substituted* I mean that one can read articles and book chapters in which design and a data collection instrument are used interchangeably, for example survey (a *design*) and questionnaire (a *data collection instrument*).

Data has to be collected, analyzed, and interpreted. Designs, on the other hand, are identified and implemented. If we overemphasize data and underemphasize design in a research project, we run the risk of collecting our data without being informed by the insights the design offers. For example, if we use a questionnaire as a primary data collection instrument without first thinking about survey design, we might jump into administering the questionnaire without thinking about the type of sampling we want to do or to what population it will or will not generalize. Doing these things after we have collected the data is often too late.

Language teachers generally feel more comfortable with data and less so with design. This is true for several reasons: First, as consumers of research, we are accustomed to seeing data and data instruments. Over the years, we have become familiar with data collection instruments like questionnaires and interviews. Another reason we feel more comfortable with data than design is that we have more often read and taken courses on DCIs than on design. As a result, data seems clear and obvious, while design seems hidden and difficult to understand.

There is a tendency to link or associate certain data collection instruments with certain research designs. Thus, questionnaires are associated with survey design to the extent that the terms are sometimes used interchangeably. Similarly, observation is associated with case study design and tests are associated with experimental design. A working tenet of this text is that while recognizing the historical relationship between particular instruments and designs, any number of data collection instruments can be used to implement any research design.

Triangulation of data

Triangulation is generally defined as a combination of methodologies in the study of the same program (Patton, 1990, p. 187). Researchers use triangulation to validate data collection instruments and meet the charge of subjective bias from single-methods or single-observer studies (Allwright & Bailey, 1991, p. 73; Jick, 1979; Lancy, 1992; Sechrest & Sidani, 1995; van Lier, 1988, p. 13). Patton (1990) summarizes triangulation by saying:

> There are basically four kinds of triangulation that contribute to verification and validation of qualitative analysis: (1) checking out the consistency of findings generated by different data-collection methods, that is methods triangulation; (2) checking out the consistency of different data sources within the same method, that is, triangulation of sources; (3) using multiple analysts to review findings, that is, analyst triangulation; and (4) using multiple perspectives or theories to interpret the data, that is, theory/perspective triangulation. (p. 464)

The basis for the triangulation metaphor comes from navigation. To understand the navigational metaphor, imagine you are in a boat on a lake, and would like to know exactly where you are. You see the old church spire on the hill in one direction, take a measurement, and draw a line on your map. You next see the town water tower in the other direction, take that measurement, and draw

a line. Where the lines cross on your map is your location. Inherent in this metaphor is the idea that the points of triangulation (the old church spire and the water tower) really exist, and that they are independent of each other.

Triangulation is, however, a metaphor with problems. Various researchers have accepted or rejected the concept of triangulation for various reasons. I will examine three positions taken towards triangulation, two that reject it and one that accepts it. The rejection of the triangulation metaphor by Guba and Lincoln (1989) is on ideological grounds; its rejection by McFee (1992) is on logical grounds; and its acceptance by Patton (1990) is on pragmatic grounds. After a brief discussion, I offer my conclusion on an understanding of the triangulation metaphor.

Two arguments against triangulation

Guba and Lincoln (1989) reject triangulation as a credibility check because "triangulation itself carries too positivist an implication, to wit, that there exist unchanging phenomena so that triangulation can logically be a check" (p. 240). The objection is the assumption that the reference points really exist. Their objection is that a type of data does not have the independent existence that, for example, the water tower in the previous example does.

Discussing the navigation metaphor of triangulation, McFee (1992) argues that triangulation is not a useful way of thinking about research. Limiting his discussion to the first type of triangulation mentioned above (i.e., methods triangulation), McFee argues that if one thinks about triangulation between methods, one assumes that both methods investigate the same thing. This is problematic because it is not clear that theoretical perspectives can be integrated. If we think about triangulation within methods, we can be sure that the methods describe the same thing, but how can we be sure we have independent viewpoints? Lynch (1996, p. 60) summarizes McFee's argument by saying that if students, teachers, and researchers all agree, there are no different reference points to triangulate. If they disagree, there is no principled way to decide which position to accept.

One argument for triangulation

Patton (1990, p. 193) argues that purity of method is less important than the search for relevant and useful information. Against the argument that one cannot mix methods, that is, the researcher cannot test predetermined hypotheses and still remain open to what emerges from open-ended phenomenological observation, Patton (1990) responds:

> [I]n practice, human reasoning is sufficiently complex and flexible that it is possible to research predetermined questions and test hypotheses about certain aspects of a program while being quite open and naturalistic in pursuing other aspects of a program. (p. 194)

What is one to think?

Essentially, the argument is over, and triangulation won. I agree with Dörnyei (2007) that triangulation gave rise to the fascination with mixed methods. I would also argue that triangulation can be considered desirable, at least at the level of data collection.

References for Introduction to Data Collection Instruments

Allwright, D., & Bailey, K. M. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in Applied Linguistics*. New York, NY: Oxford University Press.

Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24*, 602-611.

Lancy, D. (1992). *Qualitative research in education*. New York, NY: Longman.

Lynch, B. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University Press.

McFee, G. (1992). Triangulation in research: Two confusions. *Educational Research, 34(3)*, 215-219.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Sechrest, L., & Sidani, S. (1995). Quantitative and qualitative methods: Is there an alternative? *Evaluation and Program Planning, 18(1),* 77-87.

van Lier, L. (1988). *The classroom and the language learner: Ethnography and second-language classroom research*. London: Longman.

CHAPTER SEVEN

DATA FROM QUESTIONNAIRES

Sound questionnaire construction is a highly developed art form within the structure of scientific inquiry. (Rea & Parker, 1992)

In this chapter you will learn some key issues about questionnaires, and how to construct and validate a questionnaire for your class.

Introduction

Questionnaires are common in ordinary life. We often see them reported in newspapers and magazines, and now we encounter them on the Internet in which we are invited to "take this quiz" for the purpose of finding out if we are familiar with such things as the news events of the week, what famous people are doing, and what we and others think about world events.

Questionnaires as data-gathering instruments are popular research instruments in many fields including communication, education, psychology, and sociology. In applied linguistics, questionnaires are used not only for primary research, but also to supplement other kinds of research interests. For example, Lumley and Brown (2005) note the use of questionnaires in language testing research to gather background data on test candidates, to supply data for needs analysis, to support the development of tests for special purposes, to evaluate tests once they have been developed, and to provide information for test validation. However, when it comes to designing and validating questionnaires as research instruments, many TREEs are not as familiar with the procedures and techniques as they should be. How does one start to make a questionnaire? Is it allowable to use a questionnaire designed and used by another researcher? Is a questionnaire the same as a survey, or is a survey different from a questionnaire? (In this text, a survey is considered a type of research design and a questionnaire is considered a data collection instrument. That means a questionnaire can be part of a survey, but a survey is not the same as a questionnaire). How does one check or validate a questionnaire? These questions and their subsequent answers are important for any TREE creating a questionnaire. This chapter raises important questions about questionnaire development, supplies some answers, and points the way to further reading and investigation.

What is a general description or definition of a questionnaire? Even though almost everybody knows what a questionnaire is, it is still helpful to offer a definition. Gay and Airasian (2000) suggest a general definition of a questionnaire as several questions related to a research topic (p. 281). Even this short and simple definition is helpful as it has a heuristic function. *Heuristic* is defined by Vogt (2005) in his dictionary of statistics and methodology as something that is generally instructive. Definitions, models, and theories are heuristic when they suggest relationships, issues, and questions that we might not otherwise have noticed or thought about.

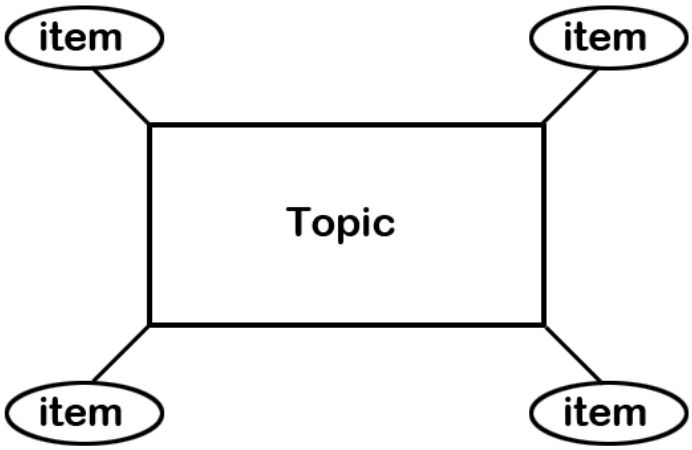


Figure 1. A visual image of the definition of a questionnaire as several questions related to a research topic

In Figure 1, we can see an image of this definition by Gay and Airasian (2000). The definition is heuristic because by showing the relationship of the definition, namely several items related to a research topic, it causes us to raise questions such as: How many items might there be? How are the items related to each other? How are the items related to the topic? What is the research topic? Where did it come from? And finally, how does the topic relate to the research purpose?

What form might a questionnaire take? There are two ways to discuss the form a questionnaire might take: administration and format. In terms of administration, a questionnaire might take the form a paper-and-pencil test, it might be administered online, it might be mailed to participants, or it might be conducted over the telephone. This chapter will concentrate on a paper-and-pencil instrument aimed primarily at intact classes.

A second way to discuss a questionnaire's form is to think about its format, that is, the physical layout on a page. Consider a pencil-and-paper questionnaire you might administer to your class. Typically, a questionnaire is formatted into three parts: 1) *demographics*, 2) *closed-ended* items, and 3) *open-ended* items.

The demographics section includes information about the questionnaire itself and information about the respondents. The purpose of providing information about the questionnaire is to identify it as a research document. Examples might include the name of the questionnaire, the version and date, and the date of administration. The second part of a questionnaire typically comprises a series of closed-ended items. The purpose of including closed-ended items is to provide quantifiable data about the construct of interest. A closed-ended item either asks the questionnaire-taker, called a *respondent*, to make a choice between two options or asks the respondent to choose an option that in some way produces a number. As a result, closed-ended

items are often easier and faster to answer than open-ended items. Examples might include true-false and Likert scale items. Questionnaires often finish with open-ended items, asking the respondent for opinions in their own words. Examples of open-ended items are short answers and sentence completion items.

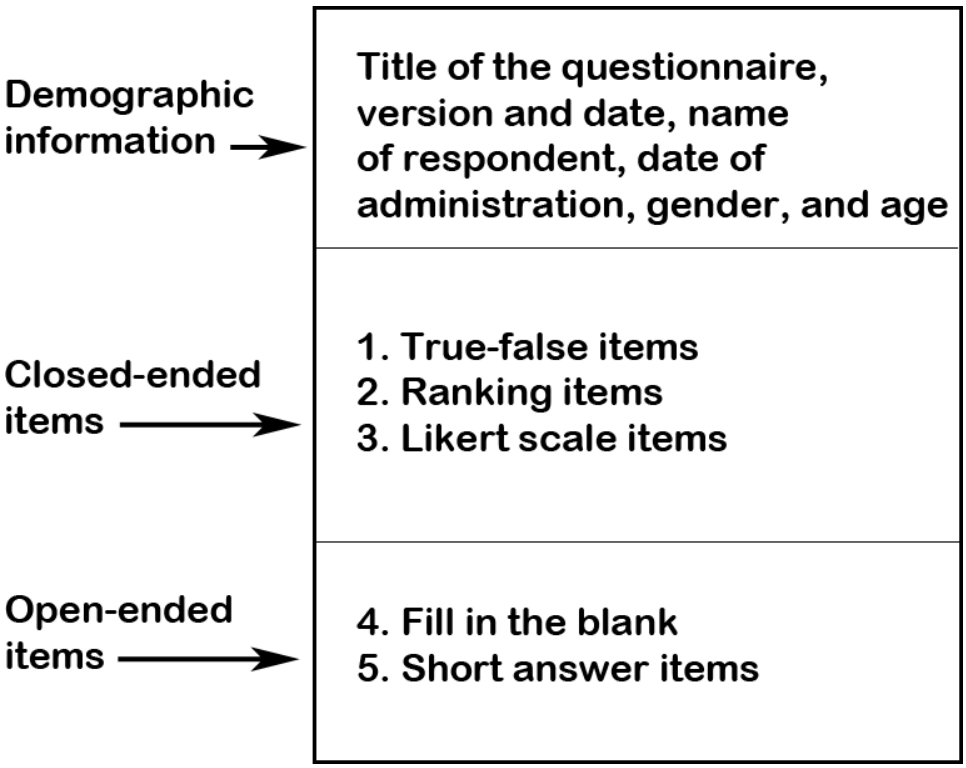


Figure 2. Layout of a one-page questionnaire showing the placement of demographic information, closed-ended items, and open-ended items

What are the advantages of questionnaires?

Listing some of the advantages (and next, some of the disadvantages) of using questionnaires is not done to recruit a questionnaire fan club, nor to discourage the use of questionnaire data. Rather, the purpose is to alert you to situations in which a questionnaire is appropriate or inappropriate, and to point out possible problems. Each advantage contains a disadvantage.

First, data from questionnaires are self-reported data. A questionnaire is an appropriate instrument for collecting data on what your students think or believe about certain issues. For this reason, a questionnaire is a standard data-gathering instrument for a needs analysis (Creswell, 2002, p. 421). But self-report data also means that respondents are reporting information, opinions, and action based on what they think, believe, or recall from previous experience. If we

want to know what people actually *do*, observation is a better form of data collection.

Second, a questionnaire is a very convenient instrument because a substantial amount of data can be gathered from a group of participants in a fairly short period of time.

Third, Creswell (2002) points out that since a questionnaire does not require the respondent's name, class members can respond to questionnaires anonymously, which might reduce the teacher influence that would be present, for instance, in an interview, where the respondent would be known.

Fourth, a questionnaire can be used to survey an entire class, group of classes, or large groups of people. If an intact class participates, it is usually possible to secure a high response rate. The disadvantage of working with large groups, for example all the writing teachers on your campus or all the writing teachers in your state, is there is typically a low response rate.

Fifth, questionnaires can be used to gather data from closed-ended items as well as data from open-ended items. Data from closed-ended items are fairly easy to collect and can be analyzed using a variety of statistical procedures. Data from open-ended items are not as easy to analyze, but using these data is still an advantage because open-ended questions provide a fuller explanation than answers that are embodied in just numbers or true/false.

Sixth, questionnaires, once developed, can be used by other researchers in different situations if the questionnaire is believed to measure the same construct. However, if you use a questionnaire developed by another TREE, the questionnaire has to be piloted and locally validated in your situation because reliability and validity typically do not transfer between populations.

Seventh, questionnaires are flexible instruments. This means that you can administer one to your class, mail it to a small group, a large group, or even administer it by telephone.

Eighth, questionnaires work well with other data collection instruments. This means they can be combined with other sources of data. For example, you can use a questionnaire to ask all members of a class if they do X, then interview selected persons to probe for details and examples of X, and then observe persons to see how or even if they are actually doing X. This process of collecting various forms of data is form of *triangulation*, which can be used to describe or possibly even explain what we are researching.

What are the disadvantages?

As mentioned previously, a low response rate is the main hazard of questionnaire research. When reading research papers and journal articles reporting the results of questionnaire data, TREEs should always look for the reported response rate. A low response rate, say below 30%, is a warning that interpretations based on the data should be regarded as tentative at best.

Second, the conditions of *self-administered* questionnaires are hard to control. A self-administered questionnaire is generally mailed to participants at their homes or offices with a cover letter asking them to fill out the questionnaire and return it by mail. A *group-administered* questionnaire, on the other hand, is a questionnaire administered in person to a group, such

as a language class, by the person who designed the questionnaire. Brown (1997) points out three problems with self-administered questionnaires: a potential for a low return rate, the need for the questionnaire to be self-explanatory, and completion at home or office under varying conditions, unknown to the researcher. If, however, the questionnaire is group administered, all these disadvantages can be overcome. By administering the questionnaire to your own class, you can have a high return rate, can clarify questions raised by participants, and know the conditions under which the questionnaire was administered.

Third, another possible disadvantage is, according to Creswell (2002): "Survey is self-reported information, reporting only what people think rather than what they do" (p. 421). This is of course true, and a good reason for having multiple sources of data in any study. If, as was mentioned earlier, we want to know what people actually do, observation may be a better form of data collection.

Fourth, data from questionnaires represents a data collection process that is considered "a mile wide and an inch deep," as opposed to interview data which might be described as "an inch wide and a mile deep." The way questionnaires are designed and administered means that usually a TREE can discuss what students answered, but can't explain why they answered as they did. For this reason, we should remember the warning from McKay (2006), "Surveys can result in superficial or simple responses" (p. 36).

Fifth, along the same line, "Surveys do not control for many variables that might explain the relationship between the independent and dependent variables" (Creswell, 2002, p. 421). One of the implications of Creswell's observation is that questionnaires do not manipulate anything; what you see is what you get.

Sixth, interpretations based on questionnaire data may not be reliable when used to gather information on actions taken (see the discussion "Do respondents tell the truth?"), but may be more reliable when gathering data on impressions, feelings, or likes and dislikes.

Seventh, a group of disadvantages surrounds the fact that questionnaires are not easy to make, but they are easy to abuse and misuse. As with any data collection instrument, questionnaire development involves a theoretical construct that must be identified, defined, and described. And like any test, a questionnaire has to be locally validated every time it is used. *Local validation* means that even though questionnaire validation information was provided when it was constructed, and even though validation information was provided when it was used previously, current validation information is still necessary.. Strictly speaking, this is not an advantage or a disadvantage, but it does mean additional work. Every time a questionnaire is administered, validation information (including reliability) has to be provided. For example, we do not know that this year's students are the same as last years, or that the Monday class responds in the same way as the Tuesday class. One reason for this is that validity does not belong to the questionnaire instrument itself, but to the interpretation based on the data derived from the instrument.

What are some key issues to consider?

Consider this imaginary but familiar scene: Courtney, an enthusiastic ESL teacher, had a concern

over the weekend about her Monday class, and she wanted to quiz her students to find out what they think about it. She wanted to know if they were reading the class textbook and the extent to which they were understanding it. Late Sunday night she sat at her kitchen table with a piece of paper and a pencil and jotted down some questions. Monday morning, she entered the questionnaire items into her computer at school and printed them out. That afternoon she administered her questionnaire, and by the end of the week she announced the "results," which impressed her colleagues since not many teachers go to the trouble of making, administering, and analyzing the results of a questionnaire. Several of her colleagues even suggested that based on the questionnaire, Courtney should submit a proposal to the next language teaching conference.

What is wrong with Courtney's questionnaire instrument and her results? What Courtney created was a series of somewhat related questionnaire items. However, she has no clear idea how her items relate to each other because she has no clear idea of what they are claiming to measure.

Maybe I am overstating the problem. She does, in fact, have an idea of what she is doing, or else why would she have done it? But the clarity she has about what she is attempting to measure is not clear to others. For example, although Courtney does not know this, at least one of her items was interpreted by some of her students to be asking if they read the textbook, and some of her students interpreted the same item to be asking if they liked the textbook. As a result, what Courtney thinks are the results of her questionnaire are seriously misleading, and the problem is that Courtney is completely unaware of this discrepancy.

Following are some of the issues Courtney could have, and in some cases definitely should have addressed in her questionnaire preparation, design, and analysis. In some cases, they are described in detail; in other cases, they are described only briefly, since they will be discussed more fully in the section titled *How can I make my own questionnaire?*

1. Construct Identified A questionnaire attempts to measure a construct (Pedhazur & Schmelkin, 1991). A construct or a concept is a purely theoretical entity, but it has certain indicators. For example, language teachers are interested in constructs such as *motivation* or *students' attitudes toward reading*. The construct the questionnaire purports to measure must be identified so readers can understand the purpose of the questionnaire. To put it another way, how can Courtney write items for a questionnaire if she is unclear about the purpose of the questionnaire (what the questionnaire is supposed to measure)? Imagine you were planning to take a trip, a friend asked you where you were going. You replied that you didn't know. If you didn't know, how would you plan for the trip, and how would you know when you had arrived? Of course, in our life we sometimes do just wander around and find interesting places, but this is not very efficient in conducting research, especially if we have to explain our procedures to others.

2. Construct sampled adequately Construct identification is related to data analysis and item writing. The issue is that all aspects of the defined construct must be represented in the questionnaire. Suppose that the construct is defined and represented by a model as having four parts. That would mean that all four parts of the model must be represented by an item or several items. If this is not done, then there is the possibility that one or more parts of the construct will not be measured.

3. Cross sectional or longitudinal Cross sectional means a questionnaire is administered only once. Longitudinal means a questionnaire is administered at least a second time, and perhaps more than that. Longitudinal research studies are especially helpful in follow-up investigations asking if certain situations are the same or have changed months, or even years, later.

4. Closed-items and Open-items As mentioned earlier, a closed-ended questionnaire item is one that gives a respondent only limited options. Examples of this are yes-no questions, with the words *yes* and *no* (or *true* and *false*) printed with instructions to circle or otherwise indicate the better answer. Another example is the Likert scale. For this, typically a statement is made with the numbers one through five labeled something like: *strongly disagree, disagree, not sure, agree,* and *strongly agree.* Respondents are requested to circle the number that corresponds with their opinion. Figure 3 is an example of a Likert scale question:

Teachers would benefit if they did research

1	2	3	4	5
strongly disagree	disagree	not sure	agree	strongly agree

Figure 3. A sample Likert scale

An open questionnaire item is one that requests the respondent to fill in a blank space or at write something, usually expressing an opinion about a topic in direct response to a question. Questionnaire designers must decide which type of items to include. Figure 4 is an example of an open-ended item:

In what ways do you think teachers would benefit from doing research?
Write your answer below.

_____

_____

Figure 4. A sample open-ended question

Closed and open items complement each other; the advantage of one is the disadvantage of the other. Closed items restrict possible responses, but that is their advantage, especially with regards to data analysis. Open items are less restrictive, but the responses are harder to code and analyze. Since Courtney, the teacher mentioned earlier, wants the opinion of her class on a curriculum idea, it would probably be a good idea to include both closed and open items in her questionnaire.

5. Rating scales (or response format) Rating scales are used to request information on a closed item. In the example used in the previous discussion of closed and open items, the scale is a Likert scale, no doubt one of the most popular scales currently used. Other scales include semantic differential scales, and true-false items. See Brown (2001) for a discussion of various scales.

6. Length of Questionnaire (how many items?) A questionnaire with many items means it is longer and takes more time to answer. The dilemma is that if there are too many items, respondents may not finish or get tired of answering questions. However, if there are too few items, we may miss getting responses about crucial issues. For a questionnaire designer such as Courtney, each item may be important for a variety of reasons. However, for the respondents, in this case her students, some questions may be interesting because they concern their class, but other items may be uninteresting, or students may not understand why they are being asked certain questions. We must always keep in mind that our respondents are doing us a favor by answering our items. Maybe Courtney's students feel they do not have a choice but to answer the questions because of their fondness for their teacher or feeling they must accept her authority. But data gathered under conditions of boredom, fatigue, or even annoyance become unreliable data. Dornyei (2003) addresses this issue and concludes that the maximum questionnaire length should be three to four pages, taking no longer to answer than 30 minutes to answer. For use in a class, my preference for questionnaire length is a maximum of 24 items and two pages; however, one page is better. Courtney may have to think about the issue of questionnaire length.

7. Piloting All data collection instruments including questionnaires must be piloted since the creator can't be sure of the respondents' interpretations of the questions. It might be helpful to do several pilot studies, each more complicated and thorough than the next. The first pilot could consist of only three or four students, who are reasonably articulate and willing to help, and who are in circumstances similar to those the target respondents will be in, (Gay & Airasian, 2000). For example, suppose you are investigating high school sophomores and juniors who came to the U.S.A. as children and were educated primarily in U.S. schools. These students are described by Harklau, Losey and Siegal (1999) as Generation 1.5, since they have language abilities somewhere between first- and second-generation immigrants. You might ask some of these students to read your questionnaire, comment on the items, explain what they think the items mean, and explain how they would answer. Their answers might provide insight into how the items are working, and might assist you in revising all or some of the items. A second pilot study might include the whole class, using the results to analyze, revise, or eliminate items. A third pilot might include several classes, and so forth.

8. Return Rate Return rate can be defined as the number of completed questionnaires the TREE receives with adequate data for analysis. Return rate is the weakness of survey design, at least in part because the TREE has little or no control over who answers and returns the questionnaire instrument. Brown (2001, p. 86) offers eight guidelines for increasing the return rate of a mail survey questionnaire:

1. Use a cover letter
2. Keep the questionnaire short

3. Offer an incentive
4. Supply return postage
5. Include a self-addressed envelope
6. Put the address on the questionnaire
7. Use follow-up letters or phone calls
8. Time the mailings carefully

Of course Courtney does not have to worry about return rate if she is only surveying her class. The problem with low return rate is that only the returned questionnaires can be counted as the sample (see Survey Research Design). Low return rate results in low reliability and low generalizability.

9. Reliability and validity An important concern, perhaps the most important concern in any data collection instrument, is how evidence of reliability and validity will be provided. Questionnaire validity concerns the relationship between the construct (what Courtney says is being measured) and the items that measure the construct (what Courtney is actually measuring). Validation is any evidence that the items are measuring the construct. Reliability is the ability of a questionnaire to produce fairly stable results. These are important issues no matter whether the data being gathered are quantitative (numbers) or qualitative (words. We suspect that Courtney is unaware of these issues, and did not consider them in her questionnaire development.

10. Respondents described When discussing survey research design, one has to make clear what (or whom) one is surveying. Almost anything can be surveyed, including abandoned cars, the number of computers in a school, students, teachers, schools, books in a library, in fact anything you can think of. Assuming one is surveying people, which in Courtney's case is her students, the total of all of the students she is interested in (would like to generalize her findings to) is called a *population* and the individual students are referred to as *units of analysis*. In the case of a questionnaire, one is restricted to people, called *respondents*. Respondents are described in terms of how the population is defined. In other words, if the population is hypothesized as comprising certain language groups, then respondents must be described in terms of those language groups. Respondents must be described so readers are convinced that participants have the same characteristics as the population of interest. If the questionnaire is being used in a survey, then it is also necessary to know how the respondents are selected.

The problem with the way Courtney conducted her questionnaire research is that she did not situate it in a research design, for example a survey research design. In the case of SRD, by not taking the survey research design into account, she did not consider design features such as population identification, probability sampling, and generalizability. If Courtney wants to use her questionnaire in another research design, she still has to consider the requirements and considerations imposed by that design. In her present situation, Courtney, by failing to designate the research design she wants to use, has conducted a questionnaire in isolation, resulting in isolated data rendering them difficult to situate and thus difficult to interpret.

11. Translation of questionnaire from TREE's L1 to student's L1 Some TREEs, especially those teaching in a foreign language situation, may wish to translate their questionnaire into the their students' language(s). Many might believe that answering questionnaire items in one's native language (L1) is less prone to misunderstanding than answering the same items in the language being studied (L2), in which respondents may have limited proficiency. I have argued (Griffee, 1999) that if teachers using a questionnaire as a data collection instrument have the questionnaire items translated from one language (for example, English) into another language (for example, Japanese), they cannot assume that the translated items are valid simply because they were translated. Valid in this situation means the items would be understood by the Japanese students in a way similar to the way intended by the questionnaire maker. Even if the original English questionnaire items in English were validated, this does not mean the Japanese questions have the same characteristics. Validity is context specific; it is not an abstract notion that transfers from one instrument to another. In other words, I argue against the assumption that a questionnaire written in one language and then translated into another language results is an equivalent instrument. Not only must the original questionnaire items be validated, but the translated items must also be validated. This is because we are often deluded into believing that since we created the items on the questionnaire and understand the meaning and intention of those items, the respondents also understand the meaning and intention of the items. To the extent that this is the case, the items are valid, and the interpretation of the results is based on solid ground. But it is possible—indeed, probable—that some students have an understanding of certain items that is different from ours.

But aren't documents, including entire books, translated and accepted as translations? Many societies acknowledge, and in some cases, revere certain translated documents. Without translation, Christians would not have access to their scriptures and the world would be without the wisdom supplied by classical Greek thinkers such as Plato and Aristotle. In the modern era, bookstores regularly sell translated documents such as philosophical essays, novels, and poetry. Even here, however, things are not as obvious as they might first appear. Miller (1992) considers translation in the sense mentioned above, and discusses four problems encountered by virtually all translators:

> 1. The syntax of one language has no equivalent in another language,
>
> 2. Words in one language don't have exact meanings in another language,
>
> 3. A word in one language has a spread of meanings that does not cover the spread of meanings in another language
>
> 4. Words that can be used figuratively in one language cannot be used figuratively in another language.

Miller concludes:

> Anyone who has translated will know the odd experience of being able to read and understand the original perfectly, as well as having native mastery of the target language, but of running constantly into unexpected and perhaps even insuperable

> difficulties in trying to turn the source text into the target language. The arrow keeps going awry and missing the target. (1992, p. 124)

Thinking more specifically about questionnaires, Widdows and Voller (1991) suggest the difficulty--if not impossibility--of a valid translation. They state, "It is interesting to note that certain concepts quite fundamental to current EFL methodology proved impossible to render into straightforward Japanese" (1991, p. 128). They add that another difficulty arises from Japanese cultural understanding of learning styles. One questionnaire item asked if the student learned better when the teacher took an interest in him or her as a person. The problem was with the word "interest" because they found "it was impossible to eradicate entirely the connotation of sexual interest in the Japanese version" (p. 128).

Yoshida (1990) conducted an experiment with second language learners (Japanese returnees who had lived in the U.S.A. for at least two years and had attended American schools). Thirty-five Japanese returnees were the experimental group, a group of 32 monolingual Japanese students in Japan were one control group, and a group of 21 monolingual American students in the U.S.A. made up another control group. All three groups were given a word association task consisting of words from nature, daily life, society, ideas, and culture. The control groups answered in their own language. The experimental group was asked to respond in Japanese to the Japanese words and in English to the English words. The two lists of words were given in different order and a week apart. Yoshida compared the responses for each word, grouped the responses into semantic categories, and calculated the degree of agreement between the experimental group and each of the control groups. His analysis showed that for the society, ideas, and the culture categories, "the bilingual group responded quite differently depending on which language they were using" (Yoshida, 1990, p. 22). For example, in giving word associations with the word *freedom*, the experimental group gave responses such as *responsibility*, *myself*, *human beings*, and *independence*, words that did not appear with the Japanese translation.

In another study, Sakamoto (1996) investigated Hyland's (1994) use of translated questionnaire items, adapted from Reid's (1987) learning style preferences questionnaire. Sakamoto's students included two groups of Japanese women aged 20 to 22 years of age at Bunka Woman's University in Tokyo. Hyland had Reid's items translated from English to Japanese, and Sakamoto used these translated items--except she retranslated four of the items she thought misleading. Sakamoto administered both the English items and the translated items to her students, allowing time between administrations to reduce the possibility of a testing effect (students remembering answers from the first test administration). She then compared the answers on the two questionnaires to determine whether the students answered the Japanese version differently than the English version. She found that "about half of the 65 participants answered the same questionnaire statements differently in English and Japanese" (Sakamoto, 1996, p. 83). Sakamoto concludes:

> Clearly, then, there were differences between the questionnaire results in English and Japanese. The high discrepancy in this study warns us that the researcher should not simply consider translation as the answer to help the respondent understand the questionnaire better. (1996, p. 87)

I am not arguing against translating questionnaire items, although the evidence raises doubts that the items in L1 will be understood and answered in the same way as the items in L2. I am, however, strongly arguing that a translated questionnaire constitutes a different instrument that, in turn, has to be subject to its own validation procedures. After piloting, analyzing, and revising items in the L2, a translated questionnaire can be administered and the results interpreted with a certain degree of confidence.

Another option would be to administer the questionnaire in L1 to students speaking a language other than the developer's L1. There are at least two instances in which this may be necessary. One is when the questionnaire is administered to a linguistically heterogeneous group of respondents, that is, a group who do not the same language. This is a common situation in North American university classes. In this case, it may not be practical to create and validate multiple questionnaires in all the target languages. Another situation that might call for administering a questionnaire in a language that is not native to the respondents happens when a researcher visits the country of the L2 students. For example, a visiting American researcher goes to Spain for a short time and does not have time to translate, pilot, and analyze items in order to create a Spanish language translation.

Validation must be built into the design of the questionnaire from the beginning (Luppescu & Day, 1990). We also know that piloting and analyzing data from the pilot must precede primary data collection. We also know that data resulting from questionnaires must be analyzed and reported. To this list, I add that we know that translation is not a short-cut validation solution. Translation results in a new document which itself must be piloted and analyzed.

12. Truthfulness of respondents A final issue is whether people tell the truth when they fill out questionnaires. Don't people sometimes, or maybe even often, lie? Of course individual persons do lie, sometimes unknowingly and sometimes knowingly, especially if they have an interest in the outcome. The issue of questionnaire research, however, is more concerned with group tendencies rather than individual ones, as individual responses tend to be "washed out" when taken as part of a group. The question then becomes, in answering questionnaires, do groups of people whom you believe to have certain knowledge, *systematically* distort what they know to be true, don't know to be true, or perhaps simply forget? If a group of math teachers are asked questions about their teaching style on one occasion, say for example how often they give lectures, how often they have students use calculators, and how often they have students work on individual projects, and then the same teachers are asked the same questions on another occasion, to what extent would the results be consistent over the two occasions? That was the question Meyer (1999) asked, and he answered his question in three ways: a composite answer, individual answers, and what might be termed quality answers based on classroom observation. Specifically, the questionnaire he administered asked selected teachers how many minutes per week they used recommended practices in their math classes.

His first answer was based on a composite of all the questions asked. On the first administration of the questionnaire, teachers reported using the recommended practices 69% of the time, and on the second occasion, they reported using them 70% of the time. The correlation between the two administrations was .69 ($p = .0013$). His second answer was based on answers to several

questions. Looking at answers to three of seven indicators, the mean scores for listening to lectures was 42.9 on the first occasion and 34.9 on the second occasion; using calculators in class was 94.0 and 87.3; and working on individual projects was 5.0 and 5.1. Expressed as correlations between the first and second administration, listening to lectures was .12, using calculators was .66, and working on individual projects was .16. These low correlations show that the math teachers were not answering consistently. Taken as a whole, the teacher reports are reliable if by reliable we mean a high correlation between first and second administrations, in other words the teachers were answering consistently, but taken on specific issues they varied considerably. For the third answer, Meyer (1999) found some teachers reported that they did not use recommended procedures, and observation verified that was the case. However, in other cases he found teachers who indicated on their questionnaires that they were using recommended practices, but in fact were not, confirmed through observation. Meyer (1999) reports the case of "Ms. Weiss," who reported that she discussed different ways to solve problems and solved problems with more than one correct answer. However, here is what Meyer (1999) observed when Ms. Weiss asked a student to explain his answer:

> Ms. Weiss chose Carlos. Would she examine Carlos's thought process? Would she help the class understand Carlos and in turn teach the class what was right or wrong in his thinking? Mr. Weiss's questions never went that deep. She practically ignored Carlos's response and just asked the class, "Does anyone disagree with Carlos or have a question about how he did it?" The class sat in silence for a moment before Ms. Weiss moved on to the next problem. (p. 42)

The questionnaire was successful at identifying responses at each end of the spectrum, namely teachers who said they did and teachers who said they did not engage in certain practices, but the questionnaire was not always successful at identifying teachers who believed they were engaging in a certain practice, but in fact were not. Do respondents lie? Probably not, but their perceptions may be different from their actions, which limits the trust we can put in questionnaire data.

How can I make my own questionnaire?

In Griffee (1997), I outline a sixteen-step process divided into five areas or stages: before writing, item writing, piloting, reliability determination, and validation evidence. An adaptation of this model can be used as a starting point in the questionnaire construction process. Creating a questionnaire might include Courtney's "Sunday night quick-and-dirty kitchen table questionnaire" or something as elaborate as a doctoral research questionnaire. The following steps should be viewed as general guidelines rather than specific steps because each step has to be evaluated and implemented in light of the particular context and circumstances of its situation. First, I list the general area and the steps (numbered in parentheses) that might be included; second, I suggest what Courtney might have done.

Before writing In the first or before-writing stage:

(1) define *the construct*, the concept you are investigating. If you are not clear what your construct is, one way to state the construct is to complete this statement: *The purpose of this questionnaire is to gather information on* _____. For many TREEs, a construct is a curious concept, and some

suspect that constructs do not exist, or if they do they are not important. They are somewhat right about the first idea (of course, constructs do not physically exist), but wrong about the second (they are not important). However, defining the construct is often easier said than done, and might be somewhat time-consuming. This is why these steps are not sequential, that is, it's rare to do one step, finish it, and then do the next. Several of these steps may happen simultaneously, or it may be more convenient to do them in a different order. You may finish a step, decide it is completed, only to find that you have to reconsider it in the light of additional experience and data. This is normal.

(2) Investigate any theory that describes the construct of interest. This presupposes, of course, that such a theory or theories exist and that you can find them.

(3) Review any previous questionnaires you can find which purport to measure the same or similar construct. Carefully examine the items in these questionnaires to see if there are any that might fit your purposes. You might also note if the questionnaire you found provides validation evidence and what type of reliability is given. Study how the evidence was gathered to determine if you want to use similar data collection and analysis procedures. Is it possible to simply use the questionnaire you found as is? If so, consider yourself lucky, and move directly to the pilot stage. But if you use a previously published questionnaire, remember to supply full citation so you are not accused of plagiarism.

(4) Decide on practical requirements for your questionnaire, such as the optimal number of items.

(5) Decide what type of data you want.

(6) Brainstorm items from yourself, others, and the literature. This means ask yourself what would be good questionnaire items,

(7) Interview colleagues and students for items. This means ask your teacher friends for good questionnaire items. For students, it means ask advanced students to make a question that would measure your construct.

Consider this stage of questionnaire development from Courtney's point of view. Courtney has an idea for a questionnaire. Maybe it's to evaluate an innovation she wants to try or has tried in her class, maybe it's to gather factual information about her students, or maybe she wants to know what problems her students have in her course and how they are dealing with them. Before Courtney sits down at her kitchen table to write her items, she might first write down the purpose of the questionnaire. Of course, the purpose of any questionnaire is to measure the construct, which is another way of saying that the purpose of a questionnaire is to measure what it is designed to measure. Writing the purpose of her questionnaire would force Courtney to state explicitly what she wants to measure, and perhaps clarify why she wants to measure it. Then she could list areas that might be relevant to her purpose, such as the sources that she might read. In addition, she might ask herself the academic keywords related to the construct she is interested in investigating: it is hard to investigate theory or review previous questionnaires if you don't know what keywords are associated with the area of interest, as used by scholars in the field.

On another piece of paper, Courtney might want to list the names of people knowledgeable in her area of interest who might be able to help her. She might locate a colleague who has previously created a questionnaire or who might be able to think of effective items. Another colleague might have ideas for questionnaire validation. Courtney can call this group of persons her "questionnaire committee." This committee might not ever meet in the same room at the same time, but identifying them and thinking of them as a group will be a helpful resource for Courtney.

Item writing In the second stage, the item writing stage:

(8) Decide how many items you want, including subtest areas. Write more items than you need so you can eliminate items that don't work.

(9) Ask your colleagues for help in item writing, if possible.

(10) Logically analyze the scoring procedures.

(11) Ask expert judges and students to review items.

Courtney is now ready to write items. She will have two major problems; first she needs to be sure each item is related to what has been referred to by various terms, such as the research topic, the construct, or the thing she is trying to measure. Second, she needs to write clear and unambiguous items that can be understood by her students. Generally, short clear items are better than long, complex items. One guideline for writing good items is to address only one issue per item. For example, the item "Do you believe in and approve of X?" are two different issues. One can believe that people have the right to do something, for example, smoke cigarettes, but still not approve of this practice. This is a good time for Courtney to read about reliability and validity as well. Courtney can ask her friends to help her write items, especially if she can explain to them the purpose of the questionnaire.

"Logically analyze your scoring procedures" is a phrase from Pedhazur and Schmelkin (1991, p. 59)' this means look critically at the relationship between the items and the construct, and determine whether the construct is adequately measured by the type of items used. For instance, if a Likert scale is used, do the Likert scale items cover the full range of the construct, or are other types of items needed, for example, open-ended or ranking items? After Courtney has a list of items, she can ask expert judges to review them. An expert judge is any person who is reasonably qualified to make a judgment about the relationship of an item to the construct. Other teachers can be expert judges. Suppose Courtney has a list of twelve possible items, and suppose she can find five expert judges. Courtney can make a form similar to the one in Figure 5.

I would appreciate your help. I want to make a questionnaire. Below are the items I want to use. Do you think my items measure my construct, that is, what I want to measure?

What I want to measure is … (insert your construct)

If you think there is a problem, make an X next to the item. We can talk about it later.

Item 1 (insert item one)

Item 2 (insert item two)

Item 3 (insert item three)

.

.

Item 12

Figure 5. Individual expert rater form

Courtney should organize all the expert results so she can understand their comments and opinions. Figure 6 shows the results of that process.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	(and so on)
Expert 1	x	x	x	x	x		x	x	
Expert 2				x					
Expert 3		x		x		x	x		
Expert 4				x			x		

Figure 6. Form showing results of all experts

Figure 6 shows what five expert judges had to say. Keep in mind that experts have personalities and backgrounds that reflect in their judgments. For example, Expert One is very critical and seems to approve of nothing. On the other hand, Expert Two is either very approving or not very critical. Variation in individual judges, their personalities and degree of criticality is not the issue. What makes a difference are the cumulative results. Item 4, and maybe Item 7, has been

flagged by these experts, regardless of their backgrounds. Courtney needs to look seriously at these items to see why the experts object. Perhaps she can ask them how she might improve the items: should they be revised or cut altogether? As far as Items 2 and 3 are concerned, Courtney will have to decide if two flags out of five warrants attention.

After Courtney reviews her items, either cutting or revising them, she can show them to students similar to the ones for whom the questionnaire is designed. She should check with all types of students, including the highest proficiency and the lowest. She should ask students, "Do you understand this question (or item as it is usually called, since not all items are questions)?" She should also ask about specific vocabulary words and grammatical constructions, and whether they are clear to the students. As I state in a 1997 publication, "It may be necessary to substitute easier vocabulary items or to paraphrase certain items, but a higher level of understanding on the part of respondents will result in less guessing, which in turn will result in higher instrument reliability" (Griffee, 1997, p. 182).

The piloting In the third stage:

(12) Pilot the questionnaire.

(13) Evaluate items.

(14) For a professional quality research questionnaire, consider a second pilot. Professional quality means writing a thesis or dissertation, or aiming a publication at a high level journal.

Courtney is now ready to pilot her questionnaire; she may wonder how many participants she needs for a pilot. One solution is to use her class, and maybe a colleague's class if she wants a larger N-size. (Recall that N means *number*, and N-size means the number of students/participants/respondents in her study).

After gathering pilot data, Courtney has to decide how to evaluate her results. There is no prescribed "this-is-the-way-to-do-it" method or methods, but Courtney has resources to draw on. She found and read several studies that used questionnaires as data collection instruments; she can now reread them, looking at how their pilot results were analyzed. She also formed a small committee to help her, and can consult with them. If she has closed-ended items, there are two possible analysis strategies. The first is, before piloting, to arrange either all or some selected key items in pairs. In other words, write two versions of the same item, administer the questionnaire containing the two versions, and then correlate the answers to the two versions. If the correlation is high, she can argue that the items are working, and she can keep one or both for her final questionnaire. The term *working* in this situation means that her questionnaire respondents (her students in this case) are answering the two items in the same way. In other words, they understand the meaning to be about the same. A low correlation indicates that respondents think the two items (both written by Courtney and both intended to mean the same thing) are being understood as different. However, in research, every solution often raises its own problem. If the solution is a "high" correlation, the problem is, what does high mean? In the final analysis, Courtney along with her advisors, will have to decide what high means, but a typical guideline is that 0.3 is at the threshold, 0.5 or higher is adequate, and 0.7 or higher is high.

The strength of each correlation will show the extent to which each item is contributing to the questionnaire. Items with low correlation are candidates for elimination. Reporting this process provides validation evidence. If this discussion of correlation doesn't make sense to you, it is an indication that you need to find a statistical consultant, learn some statistics on your own, or enroll in a statistics class.

Constructing a questionnaire is often seen as one thing and validation as another, but this is not exactly the case. For example, Courtney pilots her questionnaire items in order to improve her items. This process can be seen as a validation effort. If Courtney decides to engage in research and use her questionnaire data, she can report pilot data as validation evidence.

In the fourth stage: (15) calculate reliability and in the fifth stage, (16) explore additional types of validation. If Courtney wants to generalize her findings, she should read Chapter 3 on survey research design, because she will need probabilistic sampling to generalize her findings from her pilot to a population.

How is questionnaire data typically analyzed? Demographic data typically consist of biographical information such as gender, academic major, first language, and so on. This type of information is often frequency data, and is tabulated so that the TREE can report, for example, the number of males and females, number of chemistry majors, and the number of Chinese speakers.

Data from closed-ended items are usually drawn from a scale that quantifies the data. Depending on the scale used, this can mean frequency data (how many or how often), dichotomous data (true or false), ordinal data (ranking), or continuous data (from Likert scales, for example) often on a one-to-five scale. Numerical data can be analyzed statistically to show trends or patterns ranging from simple percentages, descriptive statistics including mean, standard deviation, median, mode, and range, to correlation, finally to more sophisticated types of analyses such as factor analysis or multiple regressions.

Data from open-ended items are qualitative (words). McKay (2006) offers a five-step process to compile and summarize open-ended item data: First, transcribe the data, probably into a document for easy manipulation. Second, think about how you intend to use the data, and group them accordingly. For example, if you posed your open-ended items to investigate X, Y, and Z, then group your transcribed data responses into three groups named X, Y, and Z. Third, read your grouped, looking for key ideas. Fourth, read your key ideas again to see if you can identify what McKay (2006) calls "reoccurring themes." These themes are a way to summarize your data. Fifth, for each theme, select a response that exemplifies the theme.

How can I calculate reliability?

As previously mentioned, estimating and reporting reliability for all data collection instruments, including questionnaires, is necessary. McKay (2006, p. 41) briefly describes three ways to do this: test-retest, repeated forms, and internal consistency. Brown (2001) has a more extensive discussion of reliability, including the formula for Cronbach alpha, the most common form of reliability reporting. Thompson (2003) edited a book that includes sixteen articles on contemporary thinking on reliability.

How can I validate a questionnaire?

Luppescu and Day (1990) offer a compelling but cautionary tale on questionnaire validation. They wanted to investigate attitudes of teachers and students in Japan about teaching English. More specifically, they wanted to develop a questionnaire that would measure what teachers and students thought were critical factors in teaching and learning. The questionnaire was to be developed to allow teachers to give it to their students to see where they and their students agreed. The assumption was that areas of agreement between teachers and students would better facilitate teaching and learning. To create such a questionnaire, they set out to define a construct they called *orientation*. Orientation was taken to mean one's orientation toward traditional teaching as opposed to a more contemporary understanding. To define orientation, they created a list of traditional teaching assumptions (e.g., L1 is the medium of instruction) and another list of contemporary teaching assumptions (e.g., L2 is the medium of instruction). They created a questionnaire of 77 items in 9 categories and administered it to 31 teachers and 84 high school and university students.

Luppescu and Day (1990) reasoned that the two lists would correlate negatively. That is, if teachers or students agreed that the medium of instruction should be L2, they would naturally tend to disagree that the medium of instruction should be L1. They found that for the teachers this was true, but for the students, it was not. In other words, the questionnaire was valid for the teachers, but not for the students. What went wrong? In analyzing their results, they concluded that they planned their study without any concern for the validation of their construct. Luppescu and Day (1990) concluded that:

> The thrust of our questionnaire was unfocused, and could best be described as a 'shotgun approach': shoot at a wide array of targets and hope that you hit something. This seems to be a common approach to questionnaire based studies. (p. 131)

In Griffee (2001) I outline seven steps to questionnaire validation that can be used as a framework; First, state the purpose of your questionnaire. It is difficult to judge the validity of a questionnaire if readers do not know its purpose. Luppescu and Day (1990) seem to do this adequately. Next, describe the construct in as much detail as possible. In the case of Luppescu and Day (1990), their construct was "orientation," and there seems to be some confusion as to whether they believed their construct to be *unidimensional* or *multidimensional*. If your construct has more than one part or is multidimensional, define each part. Third, write items that reflect the construct. You should be able to directly relate each item to your construct. If the construct is multidimensional, say it has three parts, we would expect your questionnaire to have three parts, called *subtests*, each directly matching a part in the questionnaire. On this issue, Luppescu and Day (1990) say that they failed, and they point to an item stating, "Having a conversation in English with native speakers" as neither representing a traditional nor a contemporary belief. In fact, writing items that do not fit the construct is more usual than unusual. This is because when a TREE writes an item, the item is connected to the construct in the TREE's mind. It is only later when the passage of time has weakened the link between the item and mind, or when an external reviewer points out the lack of connection, that the TREE can perceive this disconnect. The process of asking a

panel of external reviewers to judge items can be one part of the validation process to address these issues. Although they do not explicitly say so, apparently Luppescu and Day (1990) did not do this, and this is at least part of what they mean when they say that they did not validate their questionnaire. Fourth, pilot the questionnaire to see which items are working and which are not. Since Luppescu and Day (1990) did not discuss this, perhaps they did not pilot the questionnaire. Not piloting a questionnaire is a major validation omission because TREEs cannot understand how their respondents understand items until they are piloted and results are analyzed. That the items make sense to the TREE does not mean that they will make the same sense to questionnaire takers. One way the pilot can provide data to alert the questionnaire maker is discussed next. Fifth, provide descriptive statistics and a reliability coefficient. If open-ended questions were used and qualitative data were collected, describe how the data were analyzed, and what kind of reliability was used. A low reliability coefficient in a subtest or for the overall questionnaire alerts the questionnaire maker that questionnaire takers are answering certain items in a random and haphazard way. Sixth, describe the questionnaire designer (yourself, others who had a hand in writing items), your target participants, and your relationship to each other. As I wrote in 2001:

> We would like to know the number of respondents, how they were selected, and enough description (e.g., age, gender, language proficiency, school affiliation, or other relevant categories) so that the reader can grasp who these people are. This is a necessary step because it goes to generalizability, the ability of the reader to interpret the results and apply those results to his or her own situation. (Griffee, 2001, p. 11)

During their post-validation analysis of their respondents, Luppescu and Day (1990) found the heart of their validation problem. The questionnaire was designed to measure student attitudes about pedagogical ideas or assumptions, such as in which language should they study English--English, their L2, or Japanese, their L1. This presupposes, however, that the students had pedagogical ideas or assumptions about which they could have attitudes. What Luppescu and Day (1990) finally concluded was that the students, unlike the teachers, had no ideas or assumptions about language learning and teaching. Therefore, students answered all the items, both those aimed at a traditional understanding and those aimed at contemporary understanding, in what appeared to be a random, ad hoc fashion. The teachers had the construct, but the students did not. The moral of the story is don't ask questions respondents cannot answer. To put it another way, don't ask questions about a construct respondents do not have. The correct time to determine whether respondents have enough awareness of a construct to answer a questionnaire item is before the questionnaire is administered. The seventh and final step in questionnaire validation discussed here provides the actual questionnaire instrument so readers can see and inspect the items, and perhaps use them for their research purposes.

Where can I read and find out more about questionnaires?

Griffee (1997) is a rather complete journal article on validating a questionnaire. Griffee (2001), on the other hand, is a short two-page article describing what to look for in questionnaire validation. Most general textbooks on educational research methods have chapters or at least sections that discuss questionnaires. There are books on the subject of survey design and questionnaire

development. Brown (2001) is recommended for anyone interested in survey design and questionnaires; it includes chapters on quantitative data analysis and qualitative data analysis. Converse and Presser (1986) is more dated, and short (80 pages), but is still informative on the issue of writing survey questions. Dörnyei (2003) is helpful because it is short and easy to read, yet systematically covers most, if not all issues. It includes a list of published L2 questionnaires.

DISCUSSION QUESTIONS

Write some questions you had while reading about questionnaires.

Task 1. What are some issues or concerns you can identify about your school, classes, students, or colleagues that could be answered by questionnaire data?

_____

_____

Task 2. There are many issues and concerns involved in constructing a questionnaire, such as stating its purpose, defining a construct, deciding on the participants, constructing a pilot study, and validating the questionnaire, including selecting a form of reliability to report. Which could you begin on your own and which would you need help beginning?

I know how to begin	**I would need help**
1. _____	1. _____
2. _____	2. _____
3. _____	3. _____

Task 3. Looking at your answers in Task 2, interview some colleagues. Could they make any suggestions as to how you could find help?

References for Data from Questionnaires

Adler, P. A., & Adler, P. (1998). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.). *Collecting and interpreting qualitative materials* (pp. 79-109). Thousand Oaks, CA: Sage.

Allwright, D. (1988). *Observation in the language classroom*. London: Longman.

Allwright, D., & Bailey, K. M. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Bailey, K. M. (1990). The use of diary studies in teacher education programs. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 215-226). Cambridge: Cambridge University Press.

Bartlett, L. (1990). Teacher development through reflective teaching. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 202-214). Cambridge: Cambridge University Press.

Bernard, H. R. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). Walnut Creek, CA: Altamira Press.

Burns, A. (1999). *Collaborative action research for English language teachers*. Cambridge: Cambridge University Press.

Chaudron, C. (1988). *Second language classrooms*. Cambridge: Cambridge University Press.

Chaudron, C. (1991). Validation in second language classroom research: The role of observation. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith, & M. Swain (Eds.). *Foreign/ Second language pedagogy research* (pp. 187-196). Clevedon: Multilingual Matters.

Cook, V. (1989). The I-language approach and classroom observation. In C. Brumfit & R. Mitchell (Eds.). *Research in the language classroom* (pp. 71-77). London: Modern English Publications in association with The British Council.

Cosh, J. (1999). Peer observation: A reflective model. *English Language Teaching Journal 53*(1), 22-27.

Day, R. R. (1990). Teacher observation in second language teacher education. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 43-61). Cambridge: Cambridge University Press.

Evertson, C. M., & Green, J. L. (1986). Observation as inquiry and method. In M. Wittrock (Ed.), *Handbook on research and teaching* (pp. 162-213). New York, NY: MacMillan.

Fanselow, J. F. (1987). *Breaking rules: Generating and exploring alternatives in language teaching*. New York, NY: Longman.

Fanselow, J. F. (1990). "Let's see": Contrasting conversations about teaching. In J. C. Richards &

D. Nunan (Eds.). *Second language teacher education* (pp. 182-199). Cambridge: Cambridge University Press.

Fradd, S. H., & McGee, P. L. (1994). *Instructional assessment: An integrative approach to evaluating student performance*. Reading, MA: Addison-Wesley.

Galton, M. (1995). Classroom observation. In L. W. Anderson (Ed.). *International encyclopedia of teaching and teacher education* (2nd ed.). (pp. 501-506. New York, NY: Pergamon.

Gebhard, J. G. (1999). Seeing teaching differently through observation. In J. G. Gebhard & R. Operand (Eds.). *Language teaching awareness: A guide to exploring beliefs and practices* (pp. 35-58). Cambridge: Cambridge University Press.

Gebhard, J. G., Hashimoto, M., Joe, J., & Lee, H. (1999). Microteaching and self-observation: Experience in a preservice teacher education program. In J. G. Gebhard & R. Oprandy (Eds.). *Language teaching awareness: A guide to exploring beliefs and practices* (pp. 172-194). Cambridge: Cambridge University Press.

Gebhard, J. G., & Ueda-Motonaga, A. (1992). The power of observation: "Make a wish, make a dream, imagine all the possibilities!" In D. Nunan (Ed.). *Collaborative language learning and teaching* (pp. 179-191). Cambridge: Cambridge University Press.

Genesee, F., & Upshur, J. A. (1996), *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hitchcock, G., & Hughes, D. (1995). *Research and the teacher: A qualitative introduction to school-based research* (2nd ed.). New York, NY: Routledge.

Long, M. H. (1980). Inside the "Black Box": Methodological issues in classroom research on language learning. *Language Learning, 30(1),* 1-42.

Lynch, B. (1996). *Language program evaluation*. Cambridge: Cambridge University Press.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

McKay, S. L. (2006). Researching second language classrooms. Mahwah, NJ: Lawrence Erlbaum.

Richards, J. C. (1998). Through other eyes: Revisiting classroom observation. In *Beyond training*. Cambridge: Cambridge University Press.

Rossi, P. H., Freeman, H. W., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.).

Newbury Park, CA: Sage.

Spada, N. (1990). Observing classroom behaviours and learning outcomes in different second language programs. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education*. Cambridge: Cambridge University Press.

Spada, N., & Lyster, R. (1997). Macroscopic and microscopic views of L2 classrooms. *TESOL Quarterly 31(4),* 787-792.

Stern, H. H. (1989). Seeing the wood and the trees: Some thoughts on language teaching analysis. In R. K. Johnson (Ed.). *The second language curriculum*. Cambridge: Cambridge University Press.

Vierra, A., & Pollock, J. (1992). *Reading educational research*. Scottsdale, AZ: Gorsuch Scarisbrick.

Wang, Q., & Seth, N. (1998). Self-development through classroom observation: Changing perceptions in China. *English Language Teaching Journal, 52(3),* 205-213.

Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford: Blackwell.

Williams, M. (1989). A developmental view of classroom observations. *English Language Teaching Journal 43(2)*, 85-91.

Woods, D. (1996). *Teacher cognition in language teaching*. Cambridge: Cambridge University Press.

CHAPTER EIGHT

DATA FROM INTERVIEWS

Each of us has a story to tell if the right person happens to come along to ask. (Wolcott, 1995, p. 249)

What is a general description or definition of an interview?

Nunan (1992) defines an interview as "the elicitation of data by one person from another through person-to-person encounters" (p. 231). Kvale (1996) says, "[A]n interview is a conversation that has a structure and purpose," (p. 6), and Cohen, Manion and Morrison (2000) remind us that an interview is a "social, interpersonal encounter, not merely a data collection exercise" (p. 279). From this, we can conclude that as a research tool, an interview has structure, purpose, and form, and can be defined (usually) as a person-to-person structured conversation for the purpose of finding and/or creating meaningful data which has to be collected, analyzed, and validated.

How many kinds of interviews are there?

Interview types lie on a continuum between open and closed. Open interviews come from the work of Sigmund Freud, in which a patient came seeking help, was put in a controlled environment of office and couch, and was encouraged to talk about difficult situations in the hope of allowing a degree of awareness which facilitated change (Cohen, Manion, & Morrison, 2000). This type of therapeutic interview was modified to better conform to the needs of research, in which the respondent was selected by the interviewer rather than the other way around, the interviewer may have done previous investigation of the situation, and the respondent is judged to have experience helpful to the interviewer. An open interview allows the respondent wide latitude in how to answer, and may encourage the interviewer to ask probing follow-up questions. An example of an open question is: *What do you think of X?* Open interviews are often used when evaluators do not have a clear idea of what they are looking for, or when exploration is desired, as in an ethnographic study. A closed interview asks all respondents the same questions, in the same order, using the same words. In its extreme form, a closed interview asks only for factual information. Closed interviews tend to be used when the evaluators have working hypotheses they want to confirm, when multiple interviewers are employed to interview many persons, and when those answers will be compared in hopes of reaching generalizable results.

Hitchcock and Hughes (1995, p. 153) list eight interview types and two main categories they call standard interviews and non-standard interviews. Standard interviews consist of structured, semi-structured, and group, while non-standard interviews consist of group, ethnographic, life history, informal, and conversation/eavesdropping interviews.

Standard interviews In the standard category, a structured interview means questions are predetermined, the interviewer does not deviate from those questions, and the interviewer does

not ask for clarification of answers. Structured interviews are often used in market research when a fairly large number of people are asked the same questions. A semi-structured interview means questions are predetermined, but the interviewer is free to ask for clarification and even add follow up questions. Group interviews, either structured or semi-structured, are sometimes known as focus groups, and are often used to gather opinions about advertising, political positions, or marketing research.

Non-Standard interviews In the non-standard category, a group interview entails interviewing several people at the same time, for example, students who have a common characteristic, such as leadership, failure, or academic problems. An ethnographic interview is a one-on-one interview with no set agenda. The interviewer wants to know how the respondent experiences life. Ethnographic interviews have been used to explore the working life of cocktail waitresses, gang members, or school principals. A life history attempts to recreate the life of an individual. For example, a researcher may be interested in how an individual makes the transition from growing up in one country speaking one language to moving to another country and learning to speak a second language. The individual has already accomplished this feat and the researcher conducts a life history to study how it was done. Informal conversations or even eavesdropping are not really interviews, but consist of the researcher carefully listening and writing down what was said as soon as possible after hearing the information. Hammersley (1998, p. 123) describes his own ethnographic study based partially on data from conversations he overheard in a teacher's staff room. He was studying the relationship between teachers and students, and wanted to know how teachers discussed students. Listening to teachers talk among themselves in the teacher's room was an appropriate method.

Any of these interview types can be used for research purposes, but each type has its own process. Which type to use depends partly on the purpose as well as the skill and knowledge of the researcher. Probably the most common interview type in educational research is the standard semi-structured interview, because it combines predetermined questions with the ability to follow up on leads and investigate insights when they occur in the interview. The remainder of this chapter assumes you have selected some variation of the semi-structured interview.

What are the advantages of interviewing?

Many teachers are drawn to interviewing as a way of collecting evaluation data because they assume interviewing is easy to do and user friendly. Interviewing may be considered easy because it is "just talking," and talking is natural since most of us talk regularly (Griffee, 2005; Lazaraton, 2002). Interviewing may also be considered user friendly because it does not presuppose any statistical analysis, which for instructors without statistical training could be an advantage. In addition, individuals who might be candidates for interviews, such as students and other teachers, are often available and willing to talk. Finally, interview data are friendly in the sense that they can be used in conjunction with other kinds of data. For example, data from interviews can be combined with data from questionnaires to explain or strengthen interpretations. Combining types of data tends to strengthen interpretations based on that data.

What are the weaknesses of interview data?

Even though interviewing remains a popular way to gather evaluation data, Flinders (1997) offers several limitations to this method. First, individuals being interviewed, called respondents, may be inarticulate in how they understand the world. Their understandings could be implicit and taken for granted, which may mean that they may have an opinion, but they are not be able to state it in a clear way. As a result, your respondents may or may not be able to tell you what they think about the topic you are interested in. Second, some people may not have an opinion. See Luppescu and Day (1990) for an example applied to questionnaire data. Third, what people say and what people mean may not be the same. As a result, their interview data may be garbled or misleading. Fourth, persons available for interviews may not have the information or understanding we are looking for, while persons who do have the information may not be available. Fifth, people may be unwilling to discuss what they know, especially on sensitive topics. Sixth, interviewing may require skillful questioning to find out what we want to know, and certainly requires active interpretation. The results of an interview must be considered raw data that do not "speak for themselves" and require an explicit form of interview analysis (Block, 2000). As a result of all these problems, interview interpretation requires *validation evidence,* indicating the extent to which we understood that one or more of these problems might have been present, and what we did to address them.

Issues to consider

There is little agreement among evaluators, researchers, or methodologists as to the best or correct way to conduct interviews. As a result, there are many problems to address and answer. For example, one problem is starting the interviewing process without considering the desired outcome. Another problem is hesitation or reluctance to begin interviewing, which can result in putting off the interviews altogether. The first problem can be dealt with by pre-interview planning, while the second can be dealt with by piloting the interview. However, many other issues remain. What follows is a list of many of questions facing researchers who are considering interviewing. Each question is numbered, and possible solutions are offered, bearing in mind that raising questions may be more important than any particular answer.

1. Are you sure you want to interview? Metz (2001) reports that one of the few areas of agreement among educational researchers is that "the key element, the starting point, and most important issue in developing research is the research question" (p. 13). Accordingly, you should review the purpose of your evaluation, look at your research question or questions, and determine if interview data are appropriate. If what you are interested in can be directly observed, then some form of observation might be called for instead of an interview. As Patton (1990) says, "[W]e interview people to find out from them those things we cannot directly observe" (p. 278).

2. What type of interview would best suit your purpose? Earlier, Hitchcock and Hughes (1995) were cited as discussing eight types of interviews, ranging from very structured to very open. Review that list and decide which type of interview would be best for your purposes.

3. Whom to interview? One of the first and most important steps in interviewing is deciding whom to interview. Spradley (1979, p. 46) suggests individuals with a history in the situation of

interest, who are currently in the situation, and who will allow you adequate time to interview them. For much second language research, students, teachers, and administrators fulfill these requirements. One way to decide whom to interview is to create interview criteria. For example, suppose you were interested in researching a certain sport. You could create three questions, such as: *1) Do you like this sport? 2) Have you ever played this sport? 3) Are you currently playing this sport (or attending situations where this sport is played)?* Anybody answering 'yes' to two out of three questions could be considered an informed respondent.

4. What questions should be asked? Patton (1990, p. 290) discusses questions in detail, and claims there are at least six kinds of questions we can ask in an interview:

1) Experience questions about what a person has done

2) Opinion (or value) questions that tell us what people think about an issue

3) Feeling questions that are aimed at the emotional level

4) Knowledge questions that seek to find out what people know

5) Sensory questions that seek to determine what respondents have seen, heard, touched, tasted, or smelled

6) Background questions such as age, job, residence which relate the respondent to other persons

A related issue is *how* we ask our questions. To monitor questions, I have found it helpful to transcribe the interview, paying attention to how I asked questions. In one interview, when asked to repeat a question, I noticed that I tended to explain the question in great detail, and in one case I began to answer my own question. After becoming aware of this tendency, when asked to repeat the question, I simply repeated the question exactly as before, which solved the problem.

5. When should the interview end? For an individual interview, Wolcott (1995) suggests stopping when you have the data you want. This means, of course, that you understand what you want to learn from the interview. In one study, I ended an interview when I had not only asked all the questions I set out to ask, but had also explored new issues that arose.

6. How many interviews are needed for each respondent? A single interview constitutes cross-sectional research in the sense that you have only one chance to gather information. A series of interviews constitutes longitudinal research in the sense that you gather information over time. A single interview is easier to complete and analyze, and may be adequate for your purpose. Multiple interviews, however, allow the possibility of observing development. How many interviews should one strive for? There is no definitive answer to this question, because it depends on the research purpose. In the study just mentioned, I interviewed all teachers three times because my research question was to know how teachers felt about various parts of the course over the semester. I conducted three interviews for each teacher, one before the course began, in order to judge expectations, another in the middle of the course, in order to judge how things were going, and a final interview at the end of the course. It didn't make any difference

to me what the teachers reported, I simply wanted to check with them over one semester. In another study, I interviewed one teacher multiple times until I felt I had exhausted all aspects of the question of interest.

7. How many respondents should be interviewed? Again, there is no correct answer. On one occasion, I interviewed all of the teachers who taught in a program I was researching. I also interviewed representative students, selected by proficiency level--a very high level speaker, a relatively high level speaker, a middle level speaker, and a low level speaker. These students, from various classes, represented a cross-section of all students, which is what I wanted. You should devise a rationale consistent with your research question, and which will make sense to your readers.

8. Where should the interview take place? This could be anywhere that is defensible. By defensible, I mean any location that can reasonably be explained to readers. Offices may offer privacy, but conducting interviews in private offices may also be against school policy. A pilot study may help determine the best location. (See the section on pilot studies for the pros and cons of private interviews.) The important thing in writing up your research is to state the location, and describe in enough detail that your readers can judge for themselves the role your location may have played.

9. How should you introduce yourself and begin the interview? Holstein and Gubrium (1995, p. 40) describe the beginning of what they call the "traditional interview" as 1) a brief introduction of the interviewer, 2) a short explanation of the research, and 3) a request for participation. They claim, however, that any context we give provides the context for the interview: "Starting with the very introduction of the interviewer and the study itself, the interviewer offers resources and points of reference for the conversation to come" (p 40). Their point may be that we either do not pay much attention to the context we provide, or that we are trying to get out of the way of the respondent, which they would claim is impossible. It might be a good idea to write out our introduction, and consider the context we wish to provide to our respondents.

10. How should the data be collected? The answer to this concerns a matrix of issues: should we listen and take notes later, listen and take notes on the spot, or record the interview? If an interview is recorded, should the entire interview be transcribed, should a partial transcription be done, or should the interviewer just listen to the recording? The next section on conducting a pilot study addresses this issue.

11. Will a pilot interview be conducted? There are two purposes for pilot interviews: one is to conduct a practice interview for the interviewer, and another purpose is to get feedback from the respondent, which could be negative or positive. I used a pilot interview to get started, and received both negative and positive feedback. I selected a former female student with whom I felt comfortable, but who was no longer in my class. I had reservations about interviewing a young woman alone in my office, so I met her in the school cafeteria. I also had reservations about using a tape recorder because I thought it would be too intrusive, so I planned to take notes. The negative feedback was clear and immediate. At our first meeting in the cafeteria, the noise level was so high that I could see her lips moving, but could not hear what she was saying. In

addition, since it was a public place, a friend of hers asked to join us, soon followed by a faculty friend who wanted to chat. For the next interview, I changed our interview location to my office. After two interviews during which I took notes, I decided that I was losing too much information, and for future interviews I needed to record and fully transcribe each interview. The positive feedback was that my pilot interview supplied me with a major hypothesis for the evaluation, one of which I was previously unaware, namely that students were actively discouraged from speaking in class. This was news to me because the official position of the school is to run most of its classes on a seminar style. The pilot interview not only gave me feedback about the noise level and interview location, but also supplied me with a rival hypothesis that I was eventually able to confirm in interviews with other students.

What kind of data typically result from an interview?

The most typical data are the words that convey the thoughts and ideas of the respondent. But various kinds of questions can produce other kinds of data. For example, it is possible to collect numerical data by asking a series of true-false questions, "how many" questions, or "how often" questions. It would also be able to count certain linguistic features.

How is interview data typically analyzed?

There are currently several assumptions regarding interview data. One assumption is that interview data exist in a complex relationship; they are not simply a "product" from the respondent. In other words, interview data are not just the words the respondent tells us. Imagine you are conducting an interview. If it were the case that the interview consisted of only what your respondent said, we would have to pretend that *you* did not exist. In fact you *do* exist, your assumptions exist, your biases exist, and your questions exist. What constitutes *the interview* is co-created by you and the respondent. It is not just the words coming from the respondent and recorded and transcribed.

Another assumption underlying interview data is that the words from the interview constitute raw data. Raw data alone do not tell us anything, they must be interpreted. Hitchcock and Hughes (1995) describe two strategies for analyzing interview data, including several specific steps. These strategies are: 1) Become very familiar with the data, and 2) Create meaning by using analytical categories. The first strategy, becoming familiar with the data, occurs—depending on how the data was collected—by going over notes many times, listening to recordings repeatedly, or constantly reading and rereading the interview transcripts. One creates meaning by the use of categories. There is some difference of opinion (perhaps a difference of approach would be a better way of putting it) about how these categories are best created.

One approach is to become very familiar with the data, and as a result, categories "emerge" or become apparent. As we look at the data, we begin to see that our respondent was talking about theme A, theme B, and so on. Pondering these themes, we finally come to understand (interpret) that our respondent is talking about X. In this approach, the categories are "grounded" in the data; that is, categories or themes emerge from the data and reflect the data. We don't impose our will (biases) on the data, but rather let the data speak to us.

The second approach is to create the themes and categories before the interview takes place. In discussing how to conduct good interviews, Wolcott (1995, p. 115) says that behind every question should be a hypothesis. That is, we are not just asking questions randomly, rather we have some idea of what we are asking and why we are asking it. This is especially true in the case of research and program evaluation. The more exploratory your research is, the more you hope for grounded categories to emerge from the data. The more you know what you are looking for, the more you will rely on categories chosen prior to the study. For example, in the study reported, I had previously decided that I was looking for both students' and teachers' opinions on particular categories, such as the role of the textbook in learning. In each interview, I included a question on what the respondents thought of the textbook. But even with my interest in preselected categories, additional categories that I had not anticipated and was not looking for emerged from the data. I followed up on them in subsequent interviews.

Here are some examples of data analysis, based on Miles and Huberman (1994, p. 55). One problem in interview analysis is moving from a fairly large amount of raw data (the interview transcripts) to the meaning of what has been said. This is not only a process of data analysis, but of data reduction. We want to go from pages and pages of words to what is important.

Step one. Listen to the recording and transcribe the interview.

Step two. Read the transcripts several times to familiarize yourself with what is being said.

Step three. Code the interview. Coding entails reading the transcript until certain themes become apparent. Identify each theme with a short word or phrase. This word or short phrase is the code. After you have your codes, define them so you can be consistent in coding across multiple interviews. For example, in coding a teacher's interview, I used several codes including "G" and "B" which stood for "grammar" and "block." I defined grammar as "references to grammar and syntax as a goal of the class or object of classroom teaching" and block as "any reference to what is bothering or hindering the teacher." Go through the transcript and mark or circle places in the transcript where the respondent discusses the theme, and write the code in the margin. I use colored markers so I can see the themes quickly. After doing this for the entire interview transcript, you have coded the interview transcript.

Step four. Write a summary of the coded data. For example, on a piece of paper (or word processing document) write the code, and under each code list what the respondent said. For example, under the code "grammar" I put two comments, one of which was "Grammar is the main context of the course." Under the code block I put seven comments, one of which was "Course grammar book not related to academic writing." I then had reduced several pages of transcribed interview data down to one and a half pages of comments under various codes. I also knew what I believed to be the number of comments made by the teacher under each code. So, for example, I knew that the teacher commented on grammar twice, but on blocks seven times.

Step five. Write a memo to yourself. Miles and Huberman (1994, p. 72) suggest writing yourself a memo that not only summarizes, but ties together the themes and compels you to write what you think it means. This last step was the most important, because what I wrote in the memo turned out to be what I learned from the interview.

How to calculate interview reliability

One definition of reliability is stability of data. According to LeCompte and Goetz (1982, p. 35), reliability refers to the extent to which studies can be replicated, According to Kvale (1996), reliability refers to how consistent results are. Some qualitative researchers have developed a parallel vocabulary. For example, Guba and Lincoln (1989, p. 242) adopt the term *dependability,* which they maintain is parallel to reliability. Dependability is concerned with the stability of data over time.

Interview data often take the form of words, and ideas that can be coded for content by someone called a *rater*. A second rater can look at the same content and code it. It is then possible to compare the consistency of the two raters and refer to this rater agreement as reliability. Miles and Huberman (1994, p. 64) offer the formula: reliability equals the number of rater agreements divided by that number of agreements plus the number of disagreements.

$$\text{Reliability} = \frac{\text{agreement}}{\text{agreement + disagreement}}$$

For example, in a needs analysis study, six ESL instructors in a summer program were interviewed. The researcher sought a general sense of how the ESL instructors understood students' problems. The researcher audio-taped the interview, produced transcripts, and coded places in the transcript where the researcher thought the instructors were identifying student problems. These problems, 23 in all, were then coded by the researcher as language problems, teaching problems, or cultural problems. Their transcripts were examined by a second rater who coded the same problems for the same three categories. On 18 of the problem areas the raters agreed, but on 5 they did not. According to the reliability formula supplied by Miles and Huberman (1994), the reliability was 18 divided by 18 plus 5, or 18 divided by 23 for a reliability of .78. Expressed in the earlier formula:

$$\frac{18}{18 + 25} = .78$$

The two raters discussed the five disagreements and finally agreed on three of them resulting in a formula of 21 agreements and 2 disagreements for a final reliability of .91, which is a high level of inter-rater reliability.

Sources of unreliability

Cohen, Manion, and Morrison (2000, p. 121) suppose that sources of unreliability reside in the interviewer, the questions, and the respondents. This list can be expanded to include other parts of the interviewing process such as interview location, environmental factors, status equality, length of the interview, and topic threat. These eight possible sources of unreliability are discussed here, along with possible strategies to improve reliability.

1. The interviewer The interviewer, either individually or as part of a team, comes to the interview with a set of beliefs and assumptions that has to be taken into consideration. These beliefs or

biases can cause an interviewer to notice certain data that tend to support her assumptions and disregard other data that do not. This tendency is called *interviewer bias*; the danger is that this bias may occur without the interviewer's being aware of it. Interviewer bias can affect how the interviewer asks questions, which questions she asks, and how she presents validation evidence. One way to deal with bias is to recognize that you are more than a passive question asker (Agar, 1980; Long, 2005). It is likely that you have opinions on the topic about which you are interviewing, otherwise, why would you ask? In a post-modern context, it is considered naïve to maintain that interviewer bias can--or even should--be eliminated.

Evaluators with quantitative backgrounds have traditionally viewed bias as a threat to be controlled for. Some evaluators from a qualitative research tradition also share this assumption, and want to avoid or minimize the effects of bias. However, others, such as Holstein and Gubrium (1995), view the interview respondent not as what they call a vessel waiting to be emptied of their true beliefs, but as a storyteller who cooperates with the interviewer in creating meaning. Thus, bias is not a problem, because the evaluator can use bias or point of view to assist the respondent in creating multiple stories, none of which is more true than another.

Another way to think about bias as a point of view comes from Altheide and Johnson (1998, p. 293). They believe in a view of ethnography that takes various points of view into consideration, both in terms of reporting and interpretation. The author's perspective is included in the list of points of view to include. The interviewer is just one more element in the mix, and just as we would report the gender of a respondent, so we would report our own gender (if we are the interviewer) as well as a number of other possible variables that might affect our point of view.

Reliability strategy To identify possible sources of bias, you should write down your thoughts about your topic, or if you have developed a set of questions, answer the questions. If your respondent agrees with you on all or most of what you wrote, a warning light should go off. It may be that you have subtly communicated to your respondent the answer you wanted or expected. Of course, this may not be the case. It could also be that your respondent's confirmation of your opinion is grounds for accepting your hypothesis, not rejecting it. But it is wise to heed the warning light by double-checking with others. I discuss this further in the section on validation.

2. The questions The questions themselves, depending on the type of interview, may be a source of unreliability over time because as the investigation proceeds, they may change. The interviewer, over time, may change some wording here and there, subtly refining the questions, for example, until late in the investigation, respondents are answering different questions than respondents earlier in the investigation did.

Reliability strategy Strategies for using consistent or reliable questions include writing them out and reading them to respondents while allowing for follow-up questions that are unscripted. Another strategy is to record the interview so you can verify which questions were actually asked. Kvale (1996, p. 152) suggests asking the same question in a somewhat different fashion more than once to check the consistency of answers. Another strategy is data triangulation. Triangulation is comparing at least two sources of data, using one data source to strengthen (or weaken) the other. (See Introduction to Data Collection Instruments for a discussion.) For

example, you could conduct multiple interviews with the same person, asking at least some of the same questions each time. If the respondent tended to respond in the same way, even if in a slightly different form, consistency could be argued. If interviews were conducted with multiple respondents, again assuming that at least some of the questions were the same, then similar answers could be used to strengthen the interpretation.

3. Respondents Individuals being interviewed represent several potential sources of interview unreliability. One source of unreliability, termed underrepresentation, occurs when only one or a few respondents say X, and X seems plausible to the interviewer, so X is accepted. The problem is that other respondents would not say X, but they were not interviewed. A dramatic example of this occurred when I was teaching in Japan. My supervisor returned from a trip to the United States and wanted to know why Americans did not like pickles. Curious, I asked why he thought that. My supervisor was in a hamburger shop, witnessed one person taking the pickles off his hamburger, and concluded that Americans did not like pickles. His sample of one person was extended to a population of three hundred million people.

A second reliability problem with respondents comes from the fact that the interviewer may like some respondents more than others for any number of reasons including that they are available, interesting, attractive, friendly, cooperative, or articulate. This is what Miles and Huberman (1994) call the *elite bias*, the tendency to respond to some people differently than others, and often these persons are the "elite." A third respondent problem is that sometimes respondents will lie, either intentionally because they wish to hide something or unintentionally because they are self-deceived.

Reliability strategies There are several strategies for respondent unreliability. One is to expand the number of respondents to see if the data remain constant. Another is to randomly sample different, especially potentially underrepresented populations. A third strategy is to actively seek those who disagree with the data you are receiving.

4. Interview location Interview location should be considered because this may affect interview results. For example, coffee shops have levels of noise, interviews in cars may be subject to distractions (no matter who is driving), and offices have ringing telephones or unexpected visitors. All these can affect the mood, feeling, and ability to concentrate of interviewer and respondent alike. There is no perfect place to conduct an interview, but different places may produce different responses in some way. If most interviews took place in location X under conditions Y and Z, but a few interviews took place in location A under conditions B and C, and those few interviews done in location A produced data somewhat different that produced in location X, this should cause concern.

Reliability strategies These strategies include describing the location in the write-up so readers know the conditions under which the interviews took place. If possible, also use the same or similar interview location to interview all respondents.

5. Environmental factors These issues, such as weather conditions, time of day, fatigue, and even level of light or darkness may also contribute to interview data reliability. Interviewing respondents in a bright, artificially lit room without windows may result in answers different from

those given by respondents in a darker room with windows allowing them to view the weather outside. The issue is not that some locations and environmental conditions are better than others, but that a respondent may respond differently depending on surrounding circumstances.

Reliability strategies for accounting for environmental conditions include briefly describing the interview conditions and accounting for exceptions. For example, if you interview at unusual times or conditions, evaluate the quality of the data to see if it is different from the rest of your data. Ask yourself if there changes in the interviewing environments that may account for the data variations.

6. Status inequality This can be defined as an unequal power relationship between interviewer and respondent, regardless of who has the greater power. An example of interviewer status would be a teacher interviewing a student, and an example of respondent status would be a student interviewing a teacher. Status inequality could threaten interview data reliability if the respondent of lesser status perceives the interviewer to have a specific position on the interview topic. This might cause the lower-status respondent to tell the interviewer what the respondent thinks the interviewer wants to hear. In turn, this might skew a truthful response (see Gorsuch, 2002).

Reliability strategies to deal with status inequality include not interviewing a current student from your class. Cohen, Manion, and Morrison (2000, p. 123) suggest thinking of a student interview as a gift, and being thankful for it. If the interviewer has the power status, he/she should try to ask questions in a neutral way. A second strategy would be to recruit a student of similar age and status as those being interviewed. This student would conduct the interviews of other students, and you would compare those results with the results of your own student interviews. Finally, a third strategy would be to formulate a list of results and ask similar students to express agreement or disagreement, perhaps using a Likert scale of one to five. The extent to which these students agree can be used as an indication of reliability.

7. Length of interview This is the duration of any single interview, usually measured in terms of minutes. Kvale (1996, p. 151) says that interviews are often too long, including unnecessary small talk; that they can be short if the interviewer knows what to ask. Length itself, however, is not the issue for reliability purposes. What makes interview length a possible source of unreliability is when one interview is substantially longer than the others. An usually long interview may result in variable results. There may be many reasons for one or more unusually long (or short) interviews such as a respondent loves (or hates) to talk, a respondent is more (or less) articulate on the topic than most other respondents, the interviewer is drawn to (or put off by) a respondent, unusual circumstances such as interviewing in a car during a traffic jam, or unusually comfortable (or uncomfortable) surroundings. If an interview is taking a long time, but is producing interesting data, not many interviewers will terminate it.

Reliability strategies include recording the time of all interviews so that unusually short or long times can be noticed, taken into account, and investigated. If the time of an interview with one respondent stands out, the interview data can be compared with an interview of more usual length with three possible outcomes. If the results are the same, the interviewer can claim

reliability by reason of triangulation. If results are not the same but provide new or additional insight, the interviewer can verify these results by means of additional interviews. If the results are contradictory, the interviewer can claim a new working hypothesis to investigate.

8. Topic threat This is unreliability induced by interviews about sensitive areas for respondents, or what Cohen, Manion, and Morrison (2000) call "research that might pose a significant threat to those involved" (p. 121). A threatening topic may be a topic in which the respondent:

- Knows persons involved in the project
- Stands to gain or lose by the outcome
- Is being evaluated in some way
- Could be embarrassed by discussing the issues
- Could be physically or emotionally harmed or rewarded
- Could be asked about sensitive topics such as sex, religion, politics, or personal finances
- Is asked about a topic which would be awkward to disagree with, for example, "Do you teach in a communicative way?"

Reliability strategies to deal with unreliability from threatening topics include asking open-ended questions like, "Are you familiar with X? rather than "Do you approve of X?" Additional strategies might include identifying respondent relationships with a sensitive topic and comparing answers with other persons unrelated or unaffected by the topic, asking additional respondents for whom the topic may be threatening to gage possible bias, to avoid such individuals for your interview, or to interview as many of these types of respondents as possible to make sure their views are included.

Strategies to validate interview data

One way to think of interview validation is to consider questions such as, "What if I am mistaken and I can't see my mistake, or what if I am living in my own fantasy?" Kvale (1996, p. 255) describes most interview studies as black boxes, by which he means the reader has to guess on all or most of these important points: What the social context of the interview was, what the instructions given by the evaluator to the respondents were, what questions were asked, what procedures the interviewer followed in transcribing, and how the interviews were analyzed. Kvale (1996) suggests clearly stating and describing the methods used so they become public and accessible to the scrutiny of others.

A second strategy for checking the relationship between raw data and interpretation is called "checks on the data" by Hitchcock and Hughes (1995, p. 180). They postulate two approaches to validating interview data: triangulation, already mentioned as a reliability check in regard to a discussion of questions, and re-interviewing. Re-interviewing can be thought of as a check on the connection between data and interpretation that starts with the respondent. After you summarize

and have an idea about your interpretation, take the transcript and your interpretation back to the respondent. Ask if he or she agrees with your interpretation. If the respondent does, you can argue that your interpretation is more than your opinion. If the respondent disagrees with your interpretation, discuss the differences until you understand the source of disagreement. You then have a choice of deciding whether your respondent is mistaken, or reanalyzing your data and arriving at a new interpretation that takes your respondent's insights into account.

A third validation strategy is to search for negative or contradictory evidence (Fontana & Frey, 1994, p. 372). The search for negative evidence is the intentional seeking out of malcontents to see why they did not like the innovation. Suppose you interviewed several teachers, and found they all favored a certain innovation. You could then seek other teachers opposed to this innovation; obviously, they see the situation from a different point of view. Including their point of view in your analysis strengthens your interpretation.

A fourth validation strategy involves an informed, but neutral outside colleague. It is preferable to find someone knowledgeable in the subject and/or a person who is known for being critical. Your first question would be, "Given my data, coding, summary, and interpretation, can you trace a straight line from data to interpretation?" In other words, can this individual follow your line of reasoning? If you pass, go on to the second question: given that this critical colleague can trace your line of reasoning, are your conclusions (i.e., your interpretations) plausible? You are not asking if they are "correct" or if they are "true." You are asking if they make sense. Does your critical colleague think the evidence supports your conclusions? Your third question would be to ask this critical peer whether he or she can reach an alternative interpretation based on the same evidence. If your colleague cannot, you have strengthened your argument. If he or she can, you have an opportunity (after you have stopped swearing or crying) to examine the interpretation and perhaps re-interview your respondents, looking for a deeper understanding.

Remember, interview validation is not the search for perfection, although it strengthens the case for objectivity. Interview validation means trying to address the natural tendency we all have to think that our view is the correct one. By making the implicit more explicit, we open ourselves up to the points of view of others.

Where can I read and find out more about interviewing?

There are not many sources authored by language teachers on how to conduct and analyze interviews. Mostly, we have to rely on texts in general educational or ethnographic research. Hitchcock and Hughes (1995) *Research and the Teacher: A Qualitative Introduction to School-based Research* is one such text with a chapter devoted to interviewing and another to life histories. I found the chapter on interviewing helpful.

Michael Quinn Patton (1990) *Qualitative Evaluation and Research Methods* (2nd edition) devotes Chapter 7 to qualitative interviewing, a substantial amount of space to devote to interviews. This chapter represents the traditional view of qualitative interviewing in that it does address theory or analysis much, but rather assumes a more ethnographic approach. Interview planning, however, is discussed along with types of interviews. The main strength is the amount of attention Patton gives to questions, what kinds there are, and how to ask them.

Holstein and Gubrium's (1995) *The Active Interview* presupposes a general familiarity with the subject, but it is still valuable. As part of the publisher Sage's Qualitative Research Method Series, *The Active Interview* introduces what the authors call an active interview, as opposed to a traditional interview. An active interview combines the "what" (the data, what the respondent reports) with the "how" (the role the respondent plays relative to the interview). Various roles tap into different "stocks of knowledge." In their example, a woman discusses caring for her aged mother in one context as a daughter, another as a mother, and another as a wife. The weaving together of these contexts with the interviewer allows for the co-creation of meaning in an interview.

Kvale's (1996) *Interviews: An Introduction to Qualitative Research Interviewing*, at 300+ pages, is a comprehensive, yet easy to read book on the subject of interviews. It is hard to highlight any topic over the rest. However, two sections from the table of contents merit interest: Part II on conceptualizing the research interview and Part III on the seven stages of an interview investigation. This book comprises a nice combination of theory and practice.

DISCUSSION QUESTIONS

Write some questions you had while reading about interviews.

Task 1 Below are eight types of interviews, listed from most closed to most open. From your perspective, which type would be the most appropriate for your research? Why?

A. Structured

B. Semi-structured

C. Group-formal

D. Group-problem

E. Ethnographic

F. Life history

G. Informal

H. Conversation/eavesdropping

Task 2 Using the type of interview you selected in Task 1, list the advantages and disadvantages of that type.

_____

_____

_____

_____

Task 3 Here are some steps to get started. You can use them as a checklist.

- How would you state the purpose and RQs; is interviewing appropriate?
- What interviewer bias might be present; what can be done about it?
- What type of interview will be used?
- How will respondents be selected?
- How many interviews per respondent are desirable?
- What ethical considerations would you take into account?
- Where should the interviews take place?
- How will you introduce yourself and explain your project?
- What questions will you ask? Do they all directly relate to RQs?
- How will the data be collected?
- Are you going to do a pilot interview?
- How do you plan to analyze your interview data?
- Do you have a validation plan?
- How do you plan to write your report?

References for Data from Interviews

Agar, M. H. (1980). Who are you to do this? In *The professional stranger: An informal introduction to ethnography* (pp. 41-62). New York, NY: Academic Press.

Altheide, D. L., & Johnson, J. M. 1998. Criteria for assessing interpretive validity in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.). *Collecting and interpreting qualitative materials* (pp. 283-312. Thousand Oaks, CA: Sage.

Bock, D. (2000). Problematizing interview data: Voices in the mind's machine? *TESOL Quarterly, 34(4)*, 757-763.

Books, M. (1997). In-depth interviewing as qualitative investigation. In D. T. Griffee & D. Nunan (Eds.). *Classroom teachers and classroom research* (pp. 137-146). Tokyo: Japan Association for Language Teaching.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge.

Flinders, D. J. (1997). [Review of the book *InterViews: An introduction to qualitative research interviewing*]. *Evaluation and Program Planning, 20(3)*, 287-288.

Fontana, A., & Frey, J. H. (1994). Interviewing: The art of science. In D. Denzin & Y. Lincoln (Eds.). *Handbook of qualitative research* (pp. 361-376). Thousand Oaks, CA: Sage.

Gorsuch, G. (2002). Commentary on Japanese fragments: An exploration in cultural perception and duality—Making art out of fragments: An accessible vision? *Asia Pacific Journal of Language in Education, 5(1)*, 29-36.

Griffee, D. T. (2005). Research tips: Interview data collection. *Journal of Developmental Education, 28(3)*, 36-37.

Hammersley, M. (1998). *Reading ethnographic research: A critical guide* (2nd ed.). London: Longman.

Hitchcock, G., & Hughes, D. (1995). *Research and the teacher: A qualitative introduction to school-based research* (2nd ed.). New York, NY: Routledge.

Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Thousand Oaks, CA: Sage.

Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.

Lazaraton, A. (2002). What is an interview? In *A qualitative approach to the validation of oral language tests* (pp. 38-40). Cambridge: Cambridge University Press.

LeCompte, M. D., & Goetz, J. P. (1982). Problems of reliability and validity in ethnographic research. *Review of Educational Research, 52(1)*, 31-60.

Long, M. H. (2005). Methodological issues in learner needs analysis. In M. L. Long (Ed.). *Second language needs analysis* (pp. 19-76). Cambridge: Cambridge University Press.

Luppescu, S., & Day, R. (1990). Examining attitude in teachers and students: The need to evaluate questionnaire data. *Second Language Research, 6*(2), 125-134.

Metz, M. H. (2001). Intellectual border crossing in graduate education: A report from the field. *Educational Researcher, 30*(5), 12-18.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Spradley, J. P. (1979). *The ethnographic interview*. New York, NY: Holt, Reinhart & Winston.

Wolcott, H. (1995). *The art of fieldwork*. Newbury Park, CA: Sage.

CHAPTER NINE

DATA FROM OBSERVATION

Successful observation requires something more than just sitting and watching. (Lynch, 1996, p. 108)

In this chapter you will learn the difference between ordinary observation and research observation, the advantages and disadvantages of classroom observation, what observer roles can be taken, and some observation techniques. When addressing readers, I use the term *TREE*, *she*, or *you*.

Preview Questions

1. Do you believe that "seeing is believing?" (Or, is its converse, "believing is seeing" true sometimes as well?)
2. Have you ever seen something that turned out not to be what you thought?
3. What are some aspects of the second language classroom that can be observed?
4. What are some aspects that cannot be observed?
5. Have you ever observed any aspect of teaching and learning that you thought was interesting and potentially worth researching?

Introduction

Even though observation is a basic source of human knowledge going back to Aristotle and Herodotus (Adler & Adler, 1998), ordinary observation is not always a reliable source of information. Eyewitnesses are often wrong when observing action, even at close range and in good light. Here is a well-known experiment demonstrating this point: A researcher is lecturing a class (often on the topic of observation), when suddenly two people making a lot of noise enter the room wearing dark clothes. In one scenario, a man enters the room with a knife in his hand being chased by a woman with a gun. After a brief time with much yelling, the man runs out of the room followed by the woman. Immediately, the researcher asks the class to take out a piece of paper and write down what just happened. In many cases, the gun-carrier is changed to the man, and it is the woman who is being chased. This experiment is taken to illustrate that not only are our observations not reliable, but that basic facts can be changed to fit our cultural and social biases. We often see what we expect to see.

In spite of this shortcoming, observations can be used for research data collection (Nunan, 1992), and also for teacher training (Day, 1990; Wang & Seth, 1998; Williams, 1989). However, these two uses of observation (research vs. teacher training) are not closely related (Cook, 1989, p. 72). This chapter describes observation for research data collection, drawing from four literature sources: applied linguistics, evaluation, education, and anthropology. Applied linguistics literature discusses observation from a classroom research point of view; evaluation literature

looks at educational programs, but also other types of programs such as medical, drug treatment, business training, and government programs; education literature is primarily concerned with conducting K-12 curriculum research; and anthropology observes people in their local situations.

Classroom observation defined

Ordinary observation tends to be sporadic, random, and accidental. For example, we are walking down the street and see something happening in the park across the street; we are teaching a class and notice something interesting; we walk by the open door of our colleague's classroom and hear something that gives us pause. These are all examples of observation, but they are random and unplanned. On the other hand, research observation—including but not limited to classroom research—must be systematic, intentional, and theoretical. *Systematic* means the observation is not occasional, but must be principled so that it covers the area or time of interest (Genesee & Upshur, 1996, p. 85; Hitchcock & Hughes, 1995, p. 234). *Intentional* means the observer has a reason for observing. These reasons may be specific or vague, but the observer must have something in mind when observing (Adler & Adler, 1998, p. 80). The reason or reasons for observing can, in principle, be thought of as research questions or emerging research questions. *Theoretical* means the observer is working with or looking for underlying principles. In fact, Long (1980, p. 12) asserts that an observational instrument implies a theoretical claim about second language learning and teaching. Stern (1989, p. 208) opines that by ignoring theory, we run the risk of being overwhelmed by details that we can't explain. Putting these ideas together, we can define *observation* as the systematic, intentional, and principled looking, recording, and analysis of the results of our observation for the purpose of research.

Consider again the example of the observation experiment mentioned above. Students were sitting in a class when the incident happened. Their reports were not reliable, and thus not valid, because the observers were not paying attention to the incident. Therefore, they could not have any principled awareness, much less a systematic way of observing. But let's imagine that at least one of those students is again sitting in the classroom or maybe even teaching the class. The chase happens. During the next class, it happens again. The observer thinks, "I should pay attention to this." The next time it happens, our observer takes careful notes, interviews others to see what they believe happened, and maybe prepares to video the class hoping to record the event. Now we can argue that these observations have, or are beginning to have, a fairly high degree of reliability and validation, and could constitute observation data for research purposes.

What form might classroom observation take?

Classroom observation for research generally takes three forms. First, the TREE can observe her own class or the class of another colleague. If the observation takes place in a class taught by another teacher, the observer has more time to observe and record the results of the observation. If the observation takes place in a class taught by the teacher herself, then there may be less time and opportunity to record the events using paper and pencil for notetaking. This means the physical recording of the data may occur at a later date. Second, the observation items may be open or closed. *Open* means the observation item does not specify in advance what to look at or record. *Closed* means the items are specified in advance. Open means the TREE is interested in

what is happening, but has not determined exactly what she or he is looking for. Closed means the TREE has decided what she or he is looking for. Third, the data gathered may be quantitative, for example, frequency counts, or qualitative, for example, verbal descriptions. These questions can help you decide which observation technique to use: Are you observing another teacher's class or are you observing your own class? Do you know what you are looking for (closed), or are you starting with no predetermined categories (open)? Do you prefer data in numbers or words? Table 1 shows how the answers to these questions provide eight possible categories of observation.

Table 1. Eight forms classroom observation might take

Orientation	Observation category	Collecting data consisting of Numbers	Words
Observing another teacher's class	Open	1. Things you notice you can count, but were not looking for when you started your research.	2. Initially you were not looking, but as the course went on, you decided to look at this more closely.
	Closed	3. This is structured observation, for example a checklist with predetermined categories. You observe and check the number of times you see an example of the category.	4. Noting and describing high inference categories such as, are teaching tasks communicative?
Observing your own class	Open	5. Noticing something countable or fixed in space or time. For example who comes in late and by how many minutes?	6. Observing a class with no predetermined category. For example, audio recording a class, and analyzing it later.
	Closed 7.	This involves recording your data using numbers or ticking a box or category every time you notice a predetermined event.	8. Using preset categories to observe students and recording what you observe using words. An example could be noting how students perform a certain task.

What are the advantages of observation?

First, a researcher can directly observe a program (Rossi, Freeman, & Lipsey, 1999, p. 183) which means she can focus on areas of interest including specific issues, materials, activities which can be observed and evaluated (Fradd & McGee, 1994, p. 43). For example, when I was teaching in Japan, some of the local Japanese high school teachers claimed they used English extensively in their classrooms. I observed a class and found that the only use of English was about 8 seconds when the teacher read a few sentences from the board. In addition, if one teacher is observing another teacher's class, observation data allows an understanding of the program not possible by relying only on interviews (Patton, 1987, p. 12). Observation can give a "you-are-there" point of view to readers not possible from other types of data (Patton, 1990, p. 203).

Second, a TREE can search for categories of research and evaluation that make sense to participants (Alder & Alder, 1998, p. 89), or can be more focused by using predetermined categories (Allwright & Bailey, 1991, p. 3). Predetermined categories can result in numbers that can be treated statistically and analyzed (Hitchcock & Hughes, 1995, p. 235).

Third, since observation can be done by a teacher of her own class, students are not threatened by outside observers, and can do their best work (Fradd & McGee, 1994, p. 43).

Finally, we have an advantage combined with a possible disadvantage. While observation data can be rigorous when combined with other types of data (Alder & Alder, 1998, p. 89), to understand what tallies, numbers, or checkmarks mean, it may be necessary to interview the teacher and maybe the students to understand what they tallies or checkmarks mean (Hitchcock & Hughes, 1995, p. 238; Stern, 1989, p. 211). This is because a list of numbers derived from observation do not interpret themselves. Observation may show us what is going on in a classroom, but not why.

What are the disadvantages?

There are several disadvantages to observation. First, not everything can be observed (Patton, 1987, p. 12). We can see only the surface of things (Richards, 1998, p. 142). Second, maintaining observer openness is difficult because TREEs are familiar with teaching and classrooms (Bernard, 1994, p. 149). In anthropological fieldwork, for example, often the anthropologist visits the village as an outsider and observes. Teachers, on the other hand, are not outsiders; rather, we are villagers observing our own village. Because of their familiarity with classrooms, teachers have to be trained, or train themselves, to note and question the obvious rather than take it for granted and perhaps for that reason, not see it.

Third, observer education is not easy to learn and difficult to implement. Gebhard, Hashimoto, Joe, and Lee (1999) note that in graduate level courses, when observing a class, students trained to make specific comments such as "I saw the teacher write new vocabulary words on the board," regularly fell back to using vague and general comments such as, "it was a good class." As a result, if observers are not well trained, they tend to overestimate student performance (Fradd & McGee, 1994, p. 43).

Fourth, if the observer is not the usual classroom teacher, the threat of *research effect* is always present (Allwright & Bailey, 1994, p. 3). This means that an outsider observer can cause usual classroom behavior to change. In addition, it is often hard for a teacher to find somebody to observe her class (Fanselow, 1990, p. 191).

Fifth, *observer bias* is always present, since we all see and interpret events through the lens of our own experience, assumptions, and interests. Another observer may notice something we do not, or give a different interpretation. In addition, there is a high level of dependence on observer articulation (Alder & Alder, 1998, p. 104; Allwright & Bailey, 1991, p. 3; Hitchcock & Hughes, 1995, p. 235). In other words, some teacher-observers may be quite clear and articulate about what they see whereas other teachers may see the same thing and not be able to discuss it well. Data from observation does not explain itself.

Sixth, interpretation based on observation data and analysis is not adequate for most research purposes. In estimating student improvement, observation alone is usually not sufficient, and other data collection instruments, such as tests, will be needed (Fradd & McGee, 1994, p. 216).

What are some of the key issues to consider?

One key issue is the role the TREE plays in observation. Long (1980) describes two situations an observer may take: *unstructured participant observation* and *nonparticipant observation*. In unstructured participant observation, often known as ethnography, the observer does not set out to test a particular hypotheses, but takes a regular part in the activities of the group. A nonparticipant observer, on the other hand, sits in the back of the room taking notes, and does not take part in the activities. Vierra and Pollock (1992, p. 224) also describe two roles they call *participant observation* and *nonparticipant observation*. Again, a participant observer takes an active role in the activities whereas the nonparticipant sits in the back of the room and takes notes. Patton (1990, p. 208) draws a continuum with detachment at one end and full participation at the other. I accept Patton's continuum and adopt Bernard's (1994, p. 137) categories: detached observer, participant observer, observer participant, and full participant.

Figure 1. Observation participation options from detached to involved.

Detached observation would entail either observation through a one-way mirror, or having somebody video the classroom in order to watch the video later. *Participant observation* occurs when the observer is not a student and not the teacher. The participant observer would visit

the class several times so that her presence was not disturbing, but sit at the back of the room taking notes. *Observer participation* occurs when the evaluator is much closer to the class. An example of observer participant would be where the evaluator/researcher is also the teacher. The fourth category is *full participant*, or what Bernard (1994) calls *going native*. An example of full participation would be the researcher/evaluator being a student in the same program, for example a graduate student in a seminar evaluating the seminar for the class project. Patton (1990, p. 208) says deciding which role we play should be based on what is possible and what is helpful. The point is that a TREE has several options when it comes to observer roles.

Another issue is the validation of observation techniques. TREEs in their teacher roles are accustomed to many of the observation techniques, but may forget that when using observation data for research, they must also present evidence that some care has gone into the collection and analysis of the data collected. To assist TREEs in the validation of observation data, each technique has validation suggestions.

Following are eleven observation techniques. Each is described, positive and negative aspects are noted, and a validation plan is suggested.

1. Audio Recording Audio recording is the recording of all or part of a regular class. Procedures for recording may depend on the level of technology available as well as the number of students. An audio recorder with a microphone could be placed in a central location. For example, the audio player could be placed on the floor and only the microphone attached by long cord placed on a table or stand. Alternatively, small lapel mikes (called lav mikes) can be pinned on each student or placed around their neck on lanyards. Another option is to place a small hand-held recorder in a central location (Burns, 1999; Day, 1990).

Positive aspects: While some recording equipment can be bulky, a small hand held recorder is very convenient. Once the recorder is switched on and functioning, one can concentrate on teaching, knowing that data is being collected. A recording picks up linguistic details that otherwise may go unnoticed. You can decide to listen to the recording only, or you can transcribe it, resulting in a transcription others can read and study. Because audio recording produces a permanent record, you may be able to count certain details and analyze them statistically.

Negative aspects: Audio recording may work for a class size up to 20, but what about a class size of 40? Whole class activity is easier to record than group or pair work, although this could be dealt with by providing one recorder to each group. Equipment of any kind, even a small hand held recorder, is intrusive and may disconcert students so that what is recorded does not represent typical behavior. Also, audio recording alone does not show us body language or other contextual features. Transcription is time consuming, and when listening to a recording, it can be hard to determine who is speaking. Finally, most recording devices have a limited recording range, and anything out of that range is muffled or not understandable.

Validation plan: Describe your recording schedule so the readers of your report can see how often and what days of the week you recorded. Tell them your recording procedures. What did you do first, second? Describe your equipment. Where did you put the microphone? How did your students seem to respond to being recorded? Did you tell them in advance? How did they

respond when you told them? About how many minutes did the recording last each time? What were you looking for? Is there any reason to believe that your students knew what you were looking for? If there were, you run the risk of their responses not being typical. On the other hand, maybe it didn't make any difference if they knew or not. If you think it didn't make any difference, tell us why. How did you define what you were looking for? Were the data expressed in numbers or words? If the data can be counted, what kind of reliability are you prepared to report? If you used raters, tell us how you selected them. Generally, you don't have to go into detail--just a sentence or two like, "Two fellow teachers not directly associated with this evaluation were selected as raters."

2. Checklist A checklist is a form with predetermined or closed categories, usually listed down one side of the page. Space is provided (often in little boxes) to mark the presence or absence of the predetermined category. The resulting data are frequency data. The observation task can be to check a yes/no category, or it can be to check or tick a box on a continuum (Day, 1990, p. 47; Fradd & McGee, 1994, p. 272; Genesee & Upshur, 1996, p. 86; Richards, 1998, p. 144; Rossi, et al., 1999, p. 226.

Positive aspects: A checklist is simple to make and easy to use. It can capture details of the lesson you are evaluating. *Low inference data,* meaning the action observed by the evaluator does not require much inference, can give reliable data. Data from checklists work well with other data collection methods. For example, you could audio record group work activity, and later use a checklist to collect data, providing a written record that can be easily stored for retrieval and analysis. This way, the observed data are organized and easy to analyze.

Negative aspects: A checklist works well with low inference categories but not as well with high inference categories. As mentioned, low inference means the action the evaluator is observing is clear and easy to understand, and not much inference is required (e.g., How many questions did the teacher ask?). Its opposite, high inference, means the action requires considerable interpretation (e.g., Is the native language of the students valued by the instructor?). High inference categories are hard to determine with a simple yes/no checklist item, and require corroborating data. Low inference data may be reliable, but may lack a clear connection between what is observed and what is measured. Unless a TREE records what is observed, she has no way of going back to the original situation to verify the observer's decision. In other words, readers have to accept the researcher's word. The key to using a checklist is being able to establish a clear link between the predetermined category, the observed activity, and research questions.

Validation plan: It's important to report how many categories were used, and what they were. In an appendix, you can give the actual checklist form so the reader can see how it was arranged. Tell readers why you decided to use those categories. How do you define each one? Did you show your categories and definitions to a colleague for his/her critique? How many colleagues? Did you ask them if any of the categories overlapped? How did they respond? If they offered suggestions, what did you do? How many times did you administer the checklist? It could be important to mention this, because if you administered it five or six times, it would probably give more consistent results than if you did it only once. What happened when you first used your checklist? Was it easy? Do you think you left anything out? If you think that happened, tell

your readers why. Did you conduct a pilot study with a group of students similar to those in your study to familiarize yourself with how the checklist worked? This pilot data can be reported as validation information. Piloting always strengthens your validation position. Report what you did for your pilot, what you learned, and what changes you made as a result. Were your final results supported by any other data collection techniques? If so, state which ones.

3. Teacher Hand Held Digital Recorder Hand held digital recorder (or the older technology but sometimes available tape recorder) means holding a small, battery powered recorder, and speaking directly into it. The purpose is to record observations during or soon after the time of observation. The digital or taped record can be listened to or transcribed later (Burns, 1999, p. 88).

Positive aspects: These electronic devices are available and are relatively inexpensive to buy. They are small enough to fit into pocket or bag. They use batteries, allowing freedom from cords and plugs. Some tape recorders use a standard size cassette that can be played back on a larger set when listening or transcribing. By holding it close to your mouth, the resulting recording can be loud and clear. Your recorded notes are private--for your use only, or they can be shared. Audio recording allows a permanent record to be made at the time of observation or soon after, which means the audio notes may contain details that might have been forgotten as time passed between the classroom observation and recording your observation. The recording can be played back anytime, for example, in your office, or listening in a car while driving home. You can use the recording as the basis for writing a more detailed, written record.

Negative aspects: Some teachers are not comfortable talking into a recorder, even in private. This discomfort is magnified by using the recorder in public, for example, during class, after class, or while walking down the hall. Also, sooner or later, recording equipment can fail, or batteries may die and you won't have extra batteries available.

Validation plan: The raw data is made up of the recording of yourself talking about what you observed in class. This means you have anecdotal or unanalyzed data. Your words may contain a mixture of descriptive data (what you saw) and evaluation or interpretation data (how you felt about what you saw). Behind your evaluation/interpretation lie values. To identify a value, ask yourself, "What would I have to value in order to feel about X the way I do?" Behind your values is your *teacher theory*. You can use your raw data to unravel description from interpretation, and then from the interpretation, to reveal your values and your personal theory. This would enable you to understand not only how you feel, but why.

Another validation approach is to listen to your recording and take notes, or listen and do a whole or partial transcription. A transcript provides raw data, which may be read and coded. Coding is reading the transcript and deciding what themes are present, also known as *units of analysis*. List the relevant data under each code. This process is called *data reduction* because any raw data not directly relevant to a code is disregarded. Working with your codes and data, list your interpretations. Writing up the process of going from your transcription to your interpretation allows your readers to trace your steps, and this constitutes validation evidence. Again, the idea is to document how you went from the raw data to an interpretation.

A third validation approach is to listen to your recording several times and reach conclusions. These conclusions are considered weak and unsupported because the readers have to take your word for it, and can reasonably ask, "Why we should believe you--couldn't you be mistaken?" You can strengthen your position by stating how many times you listened to the recordings, that you were able to get a peer to also listen to them, and that your peer either agreed with your interpretations or came up with similar results on her own.

4. In-Class Observation Notes In-class observation notes refer to any type of written documentation made by an observer other than the teacher while the class is meeting. Notes written after the event are referred to as a Teacher Diary. Much has been written to describe the strong and weak aspects of this type of observation documentation, which is a mainstay of ethnographic observation (Adler & Adler, 1998; Allwright & Bailey, 1991; Bernard, 1994; Evertson & Green, 1996; Fradd & McGee, 1994; Genesee & Upshur, 1996; Gebhard, 1999; Guba & Lincoln, 1989; Lynch, 1996; Patton, 1987, 1990; Rossi, et al., 1999; Vierra & Pollock, 1992).

Positive aspects: The evaluator can observe an authentic educational environment, for example, a classroom. Course goals and objectives can be verified. The observer does not have to rely solely on handouts or secondhand reports of what is happening. Specific materials, activities, and procedures can be observed and evaluated. Aspects of the course the teacher might miss, ignore, or take for granted can be observed and noted. Observation data can be words or countable items of interest, such as who speaks or where people sit. Nonverbal communication and behaviors, for example, the way people dress or the way they use physical space can be observed. Aspects of the classroom that might be meaningful, but are often overlooked--the bulletin board or room decor--might be observed and noted.

Negative aspects: Observation using in-class notes is labor intensive. In addition, not everything can be observed in a clear and obvious way. For example, high-inference categories, such as student attitudes and feelings have to be carefully defined, and what you accept as evidence carefully spelled out. Any instrument gives just one view. Observation data by itself is probably not enough, and you may also need numerical, quantitative data, or qualitative interview data. It is hard to learn how to be an objective observer and not mix descriptive observation with evaluation and opinion. There are many sources of error, and Evertson and Green (1986, p. 183) list seventeen of them, including problems with both rating scales and ethnographic observation. Another consideration is that you need to negotiate an observation schedule with the teacher that allows for observation on a systematic basis. Observing only a single class leaves you open to charges of unreliability, that is, the claim that what you observed was an unusual occurrence and only happened one time that semester, which happened to be the day you were there. If you observe more than once, but you observe on the same day of the week, you run the risk of skewing results because Friday classes may be different from Monday classes in small but important ways. You also have to decide what role you want to play while observing. For example, while observing, what would you do if the teacher or a student turned to you and asked a question?

Validation plan: If you have numerical data, see the validation plan for the observation technique called *Checklist*. Assuming your data consists of words, ask yourself if your readers can follow the actual sequence of how the data were collected, processed, and transformed for specific

conclusion drawing. Look at the validation strategy for *Teacher Hand Held Digital Recorder* to see what you can apply. Tell your readers the context of your observation: How often and when did you observe? Where did you sit? What was your level of class participation? What were the students like? What did your notes look like? How did you organize them? How did you go from observation to interpretations? What was your role in the class? In relation to the teacher, were you a peer, a supervisor, trainer, friend, or some other role?

5. Peer Observation Peer observation refers to two teachers observing each other's classes. Both teachers typically adopt a participant observation stance, which means they sit in the back of the room, observe, and take notes. A feedback session may be held later. Since they are peers, evaluation of each other is not an issue. The purpose of peer observation from a course research perspective is to gather evidence from a viewpoint not otherwise possible. Peer observation is typically discussed in the context of teacher education (see for example Cosh, 1999; Fanselow, 1990; Gebhard, et al., 1999; Gebhard & Ueda-Motonaga, 1992; Richards, 1998, p. 147; Wang & Seth, 1998; and Williams, 1989).

Positive aspects: Peer observation makes collegial research and evaluation possible. This means that two teachers working together can research the same phenomenon, but in each other's class. Also, it may be possible to use another class as a mirror to evaluate an issue. For instance, Teacher A can have an issue of interest and use observation of Teacher B's class to illuminate that issue. Teacher B may or may not even be aware of what Teacher A's issue is.

Negative aspects: Peer observation is not common, so it may be hard to find a willing partner because some teachers might see it as potentially threatening. It can also be difficult to arrange schedules. The key to participating in peer observation is to have a clear idea of your purpose, and for you and your peer to agree as to what you will do, when you will do it, and how you will share information. More than likely, your data will take the form of written notes.

Validation plan: Refer to *In-Class Observation* and *Teacher Hand Held Digital Recorder* validation plans.

6. Proformas As seen in Figure 2, a proforma is a grid or a form to fill in that can hold a type of class performance, including descriptive as well as interpretative data (Burns, 1999; Day, 1990).

Time Period	Category 1	Category 2	Category 3	Category 4

Figure 2. Example of a Proforma

Categories of interest are written across the top of a piece of paper or card. Examples of column headings could be Date, Question or area of interest, Descriptive Notes, and Interpretation. Descriptive notes can be entered during or after class. If entered after class, the proforma functions as a teacher diary, journal, or log. Interpretative comments could also be written after class.

Positive aspects: A proforma can be an all-purpose data collection instrument. Using just this one instrument, you can decide on a category, collect descriptive data, and engage in interpretation. The key is the categories, which ultimately have to be explained. Deciding which categories fit your situation may require trial and error to find the ones that give you helpful data. Questions or categories can be predetermined; this preparatory activity helps to decide what to note. Alternatively, categories can be relatively open and vague, depending on how clear you are about what you are looking for. Proformas can be put on cards for easy filing. Proformas can be taken into class, and there is space to jot notes on areas of interest. Proformas may take less time than the *Teacher Diary* because notes are shorter and more to the point than diary or log entries.

Negative aspects: One problem is that if you are observing your own class, it may be difficult to take notes while teaching. Another potential problem is forgetting to date the proforma, or making notes so cryptic you can't understand them later. Yet another problem is not interpreting your data, so that later you have data that you don't know what it means because much of the context for the data is forgotten. A potential weakness of proformas is a tendency to use subjective and ill-defined categories.

Validation plan: You should be sure to define the categories or at least explain what they mean. If you are collecting verbal, descriptive data, use validation strategies similar to the ones discussed in *In-Class Observation* and the three strategies discussed in *Hand Held Digital Recorder*.

7. Scribbles These are short and quick observation notes jotted down while teaching (Burns, 1999, p. 85).

Positive aspects: Notes made during the class have an authenticity of being made by someone who was there and experienced a behavior or activity firsthand. Scribbles can be used as the basis for recalling and writing more complete diary or proforma entries.

Negative aspects: It is difficult to teach (an active undertaking) and take notes (a reflective undertaking) at the same time. It may difficult to keep scribbles in a systematic way. If you don't use your scribbles as the basis for more complete notes right after class, they may be hard to read or understand later. Some teachers have trouble taking notes while teaching because they tend to be peripatetic, moving around the classroom as they teach. On the other hand, it might be possible to schedule brief periods of time for writing. Writing could be done in a small pocket notebook, index cards, sticky notes, or in a text file on a mobile phone, tablet, or laptop.

Validation plan: Keeping a record of each time you gather scribbles could show your observations to be intentional and systematic, two qualities that increase validity. As in *Proformas*, if you have predetermined categories, you should tell your readers what they are, why you are interested in them, how you selected them, how you define them, and how they are related to your evaluation

purpose and research questions. Probably the best idea would be to use scribbles as the basis for another observation technique such as *Teacher Diary*.

8. Seating Chart A seating chart can be made either by an observer sitting in the back of the room or by the teacher. It could be a single sheet of paper with boxes for each student showing where each student is sitting. If you allow students to sit wherever they please, you might find revealing patterns (Burns, 1999, p. 106; Day, 1990, p. 49).

Positive aspects: A seating chart could be helpful to your research if you are investigating a topic such as student social relationships. While you are making a seating chart, it would be easy to use the chart to incorporate other features of the room including the number and location of all objects, such as windows and chalkboards, and perhaps the measurements and description of the classroom. This information is helpful when it comes to writing your report because you can describe the site with more detail. A seating chart can be combined with a video record to show what happened and where.

Negative aspects: There are two kinds of data commonly provided by a seating chart: Where students are sitting and a record of what students are doing, for example, how many times each student initiates a speaking turn or other action. While sitting in the back of the room, an observer can not only make a seating chart, but also have the time to indicate student interaction. A classroom teacher can also make a seating chart, but may not be able to also conduct a class while using such a chart to record interaction.

Validation plan: To validate seating chart data, you can create seating charts on multiple occasions to show that seating patterns were consistent. Students tend to sit in the same place, but if for some reason there is no consistency in where students sit, they this observation tool should be not be used. If you draw conclusions about what certain seating patterns mean; theory, principles, or prior research would be needed for explanatory purposes.

9. Structured Observation Structured observation is classroom observation using previously defined categories. In some cases, an observation form is given to the observer with instructions to note when, how often, or examples of classroom activities that in the observer's opinion exemplify the category. This implies rater training, a theoretical basis for the categories, and instrument validation similar to that used by questionnaires and tests (Allwright, 1988; Chaudron, 1988, 1991; Cook, 1989; Fanselow, 1987; Galton, 1995; Long, 1980; Moskowitz, 1967; Nunan, 1992; Spada, 1990; Spada, & Lyster, 1997; Weir & Roberts, 1994).

Positive aspects: There are many structured observation instruments available, for example Long (1980) reviews 22 such instruments. Two of the better known are: FOCUS (Foci for Observing Communications Used in Settings) (Fanselow, 1987); and COLT (Communicative Orientation of Language Teaching) (Spada, 1990; Spada & Lyster, 1997). Despite criticism, there remains continued interest in this type of data collection. When using low inference categories, relatively objective and reliable data can be collected. Data collection may be recorded using audio or video or in real time, and resulting data can be analyzed.

Negative aspects: The main danger is lack of validation evidence, that is, the chosen categories

may exclude critical features from the data. The categories may appear relatively simple, but even seemingly simple categories are often ambiguous and capable of multiple interpretations. In addition, while one might find patterns, it does not follow that what is infrequent is insignificant. Quantification of data into numbers cannot explain what those patterns meant to the participants.

Coding categories tend to focus on what a teacher says and does, which assumes that the teacher controls classroom interaction. This may or may not be the case, for examplea structured observation instrument may not be able to capture what goes on in pair or group work. Concentrating on the teacher may also limit the usefulness of data from observation restricted in the sense that nonverbal communication is not checked. Also, raters must be trained and retrained from time to time. If an observer has a strong stake in the outcome, observer bias is possible; the presence of an observer has some effect. Although observer effect can be overcome to some extent by having neutral observers, being unobtrusive as possible, and having a sufficient number of observations.

Since the data gathered are typically a count of certain features that provide frequency data, some sort of statistical analysis and presentation is necessary. Those who aren't trained in statistics will consider this a weakness, while those who appreciate statistics will consider this a strong point. The purpose of your research is key, and any observation instrument must serve the research purpose. Much of the criticism of structured instruments comes from ethnographically-oriented researchers who point out the problems of using predetermined categories. These same criticisms can be made of any predetermined category instrument as well, such as tests or questionnaires. For example, one charge is that using a predetermined category tends to focus the observer's attention in a specific direction, and in doing so, the observer may miss something important that falls outside the category. The trouble with this criticism is that while it is true, in fact, the point of focusing your attention is precisely that, to focus your attention. If you don't know what you are looking for, you shouldn't use this type of instrument. But it is unfair to criticize a predetermined instrument for being predetermined. It is like criticizing an apple for being an apple and not an orange.

Structured observation instruments are the Rolls Royce of the observation world, and as such, they are mostly used for research. Regarding the famous car, most of us have never driven--let alone owned--one, and probably something similar can be said about using a structured observation instrument. First, you have to find or adapt the one you want to use, then, you have to be trained or train yourself to use it. Finally, you have to work out a validation plan. This type of observation instrument may require finding a mentor experienced in its use, and also requires familiarity with the literature. Be prepared to spend time in preparation, and be prepared to conduct one or more pilots, in which you not only train yourself in the use of your instrument, but also gather data for its validation. Essentially, you have three choices in using a structured observation instrument: You can make your own, you can use an existing instrument, or you can modify an existing instrument to suit your needs.

Validation plan: Before you even begin, create a log to record your procedures. Include as much detail as possible; most teachers don't gather enough of this information. The rater training should be documented in detail, including who was selected, why, and how they were trained.

This is true even if you are the rater. Be prepared to deal with the charge of observer bias, since the presence of an observer will have some effect. Inter-rater reliability is the most common as well as the recommended form of reliability measurement. Another major issue is validation of the categories. Validation involves establishing a link between the category, a theory or body of knowledge that describes and defines the category, and your research purpose. You may want to consider advanced statistical procedures such as factor analysis or structural equation modeling as well as case studies to validate your data. See Weir and Roberts (1994, p. 173), who discuss six validation strategies.

10. Teacher Diary A teacher diary, as used in this chapter, is a log or journal written primarily by a teacher after a class session is over. This document is based on observations made during class (see Bailey,1990; Bartlett, 1990, p. 209). When evaluating another teacher's class, a teacher diary serves primarily as a chronology of events and a repository for reflection. When evaluating the TREE's own class, this diary includes those two functions plus it takes the place of in-class field notes. In this way, a teacher diary functions as a source for descriptive data, and as a source for teacher reflective data.

Positive Aspects: A teacher diary is a term made popular in the applied linguistics literature for ethnographic field notes, and as such, comes from a long and accepted history. It provides a written document of what happened. If you don't write it down, your recollection of events will blur and merge with other memories until you can't remember any details; details give your report credibility.

Negative Aspects: Teacher diary data is recalled, rather than actually observed data. That means the sooner you make your diary entry after the class, the better. It is easy to put off writing (especially when you feel you have no time, tired, nothing special happened in class, or too many other things to do when you get home). It is also hard to separate descriptive from evaluative comments. Comments such as "Good class today" are evaluative rather than descriptive in that they don't tell readers what happened to make the class good. You have to train yourself to write the details of what actually happened, as well as how you feel about it. Teacher feelings are important, but they are not the same as detailing what caused them.

For a teacher evaluating another teacher's class, a diary is helpful; for a teacher evaluating her own class, a diary is indispensable. Without a diary, you run the risk of using only quantitative data gathering instruments such as tests and questionnaires, which can give reliable and valid data, but will make interpreting that data difficult. All data from your diary, even your reflections, have to be considered raw data. That means all entries have to be analyzed and interpreted.

Validation plan: If you are observing your own class, your teacher diary can become a valuable observation instrument, perhaps your only observation instrument. For that reason validation must be carefully considered. It might be helpful for you to arrange your diary into four sections: history, emotional reflection, descriptive data, and reflective data.

To offer validation evidence for your history, you should make frequent diary entries, even if you have to force yourself. At the end of the evaluation period, write a history of the course as completely as you can. Consider collecting documents, such as your local or school newspaper, to

look for external events that might have influenced your students during the observation period.

To offer validation evidence for your emotional reflection, keep a record of your high and low moods, what made you happy, depressed, elated, satisfied, and so forth. This can be an indication as to your bias, or to put it more positively, your belief system. Try to match high emotional states to events that caused or at least seem related to your high mood. Similarly, try to match events to low moods. Can you state what might be your bias? If so, you can report this as part of your validation data in that it helps your readers understand and interpret your findings. Do the same thing for your teaching beliefs. How do events and bias indicate your beliefs? Your personal theory is a filter through which all observations pass. It is helpful to report your understanding of this so readers can take it into account.

To offer validation for your descriptive data, if you used *Scribbles*, *Checklist*, or *Proforma*, use the validation strategies mentioned there. Describe in detail how you collected your data. When did you write up your observations? Take readers through your process step by step so they can follow and understand your data collection process. Search your data for themes. Code the themes, that is, give each theme a name, and see if you can relate your themes to your research questions. If you don't have satisfactory research questions yet, try to use your themes to create new ones or sharpen your present ones. Ask colleagues to look at your data, your themes, and your research questions to see if they can follow your logic. If they can, report this as validation data. If they can't, work with your themes and RQs until they can.

To offer validation for your interpretations, ask a colleague to look at your interpretations (or hypotheses). Are they plausible? If not, what additional evidence would you need? Find a critical colleague who is willing to look at your data to see if they can find fault with your interpretation and give you an "alternative hypothesis." Don't pick your best friend or someone who tends to agree with you or share all your values. Finally, see if you can match your findings (your interpretations) with findings from other sources of data, for example tests or interviews. This data triangulation strengthens your findings.

11. Video This refers to the video recording of all or a portion of a class. The purpose of video recording a class is to provide data to answer a research or evaluation question. For discussion on the use of video see Burns (1999, p. 94), Day (1990, p. 46), Galton (1995, p. 502) and Woods (1996, p. 37).

Positive Aspects: Small high-quality video cameras are now relatively inexpensive and available, and many cell phones now include a video function. Video data can reveal things we might not otherwise notice. Video can give a detailed naturalistic view of life in a classroom, and a sense of being there.

Negative Aspects: Cameras don't see everything, only what they are pointed at, and they can't be pointed at everything in a classroom all the time. You may have to buy the recording media, get the camera, and aim the camera. Video gets harder as the number of students increases. If you video a class for an hour, then you have to spend an hour watching the video, and perhaps several hours transcribing the sound track, which may not always be of high quality. In addition, video can be intrusive, and we run the risk of student reluctance to behave in a normal way in the

presence of a video camera, including students playing to the camera. There are certain ethical considerations given the fact that students may be recognizable on the tape. You may have to justify to an institutional research board (IRB) why you need to video students. If your students are under 18 years of age, videoing without parental or guardian permission is likely illegal. And yet, video is a technology that is here, can be used, and for some kinds of research, may be required. For example, if interpretation of body language is part of a research question, a visual record may be necessary. A school or department with video equipment, a budget, and staff to operate the camera would be helpful. You might want to try a pilot study to see if video is helpful, worth the trouble, and provides the kind of data you want.

Validation plan: If your data are countable, adapt the validation plan from *Audio Recording*. If your data takes the form of words, look at how this type of raw data are validated in technique number two, *Teacher Hand Held Recorder*.

In conclusion, the purpose of this discussion of observation techniques is to demonstrate that many observation techniques are available, that they can be used alone or in conjunction with others. Regardless of how they are used, they all must be validated.

Observation techniques grouped by use

In Table 2, the eleven observation techniques discussed in this chapter are grouped according to the eight categories described in Table 1. The number in parenthesis behind each technique is the number of its listing. Each of the eight groups of observation techniques is situated in a particular research situation. If you are playing an outsider role, consult sets one through four. If you are playing an insider role, evaluating or researching your own class, consult sets five through eight. Decide if you have a predetermined category (you know what you are looking for), here called *closed*, or if you do not have a predetermined category (you aren't sure what exactly you are looking for), called *open*. Decide which observation techniques fit your situation and best answer your evaluation or research questions.

Observation techniques, both in the classroom and in other research situations, are time-honored data collection instruments. However, observation of the type necessary for research and evaluation requires a certain amount of attention that goes beyond what teachers typically do. Not every observation technique fits every situation. A major difference seems to occur whether you are observing another teacher's class or your own. Sometimes validation reporting takes the form of counting and statistical analysis, but often it does not. In many classroom research studies, validation data of the data collection instruments are not reported. Whether we use quantitatively or qualitatively oriented instruments, we are obligated to report how we got the data, how the data were analyzed, how the data were interpreted, and what steps we took to investigate the integrity of our data-gathering process.

Table 2. Placement of Observation Techniques in Eight Categories

Use	Observation category	Data collected in terms of Numbers	Words
Observing another teacher's class	Open	1. Seating Chart (8) Peer Observation (5)	2. In-Class Observation (4)
	Closed	3. Checklist (2) Seating Chart 8) Structured Observation (9)	4. In-Class Observation (4) Peer Observation (5)
Observing your own class	Open	5. Audio Recording (1) Seating Chart (8) Video Recording (11)	6. Audio Recording (1) Handheld recorder (3) Proformas (6) Scribbles (7) Teacher Diary (10) Video Recording (11)
	Closed	7. Audio Recording (1) Checklist (2) Seating Chart (8) Video Recording (11)	8. Audio Recording (1) Handheld recorder (3) Proformas (6) Scribbles (7) Teacher Diary (10) Video Recording (11)

Where can I read and find out more about observation?

Brian Lynch, in his book *Language Program Evaluation* (1996), includes a section on observation as data gathering starting (p. 108). He also provides an example of a structured observation instrument, the Communicative Orientation of Language Teaching (COLT) scheme. He describes fieldnotes in detail, with a helpful discussion on the difference between descriptive as opposed to evaluative comments. Anne Burns, in a book on action research, *Collaborative Action Research for English Language Teachers* (1999), describes in detail many observation techniques, including a discussion of validation. Weir and Roberts in *Evaluation in ELT* (1994), discuss observation techniques. Finally, Genesee and Upshur, in *Classroom-Based Evaluation in Second Language Education* (1996), devote an entire chapter to observation and provide a framework for classroom observation.

DISCUSSION QUESTIONS

Write some questions you had while reading about observation.

Task 1. Describe a time when you observed something about your teaching or something that happened in your class that you thought was interesting.

Task 2. How might your observation be framed as a research project?

Task 3. Using the situation you first noticed in Task 1 and reframed as a research project in Task 2, what observation technique(s) described in this chapter could you use to gather research data?

Task 4. What additional data collection instruments might be helpful?

References for Data from Observation

Adler, P. A., & Adler, P. (1998). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.). *Collecting and interpreting qualitative materials* (pp. 79-109). Thousand Oaks, CA: Sage.

Allwright, D. (1988). *Observation in the language classroom*. London: Longman.

Allwright, D., & Bailey, K. M. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Bailey, K. M. (1990). The use of diary studies in teacher education programs. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 215-226). Cambridge: Cambridge University Press.

Bartlett, L. (1990). Teacher development through reflective teaching. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 202-214). Cambridge: Cambridge University Press.

Bernard, H. R. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). Walnut Creek, CA: Altamira Press.

Burns, A. (1999). *Collaborative action research for English language teachers*. Cambridge: Cambridge University Press.

Chaudron, C. (1988). *Second language classrooms*. Cambridge: Cambridge University Press.

Chaudron, C. (1991). Validation in second language classroom research: The role of observation. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith, & M. Swain (Eds.). *Foreign/ Second language pedagogy research* (pp. 187-196). Clevedon: Multilingual Matters.

Cook, V. (1989). The I-language approach and classroom observation. In C. Brumfit and R. Mitchell (Eds.). *Research in the language classroom* (pp. 71-77). London: Modern English Publications in association with the British Council.

Cosh, J. (1999). Peer observation: A reflective model. *English Language Teaching Journal 53*(1), 22-27.

Day, R. R. (1990). Teacher observation in second language teacher education. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education* (pp. 43-61). Cambridge: Cambridge University Press.

Evertson, C. M., & Green, J. L. (1986). Observation as inquiry and method. In M. Wittrock (Ed.), *Handbook on research and teaching* (pp. 162-213). New York, NY: MacMillan.

Fanselow, J. F. (1987). *Breaking rules: Generating and exploring alternatives in language teaching*. New York, NY: Longman.

Fanselow, J. F. (1990). "Let's see": Contrasting conversations about teaching. In J. C. Richards & D.

Nunan (Eds.). *Second language teacher education* (pp. 182-199). Cambridge: Cambridge University Press.

Fradd, S. H., & McGee, P. L. (1994). *Instructional assessment: An integrative approach to evaluating student performance*. Reading, MA: Addison-Wesley.

Galton, M. (1995). Classroom observation. In L. W. Anderson (Ed.). *International encyclopedia of teaching and teacher education* (2nd ed.). (pp. 501-506). New York, NY: Pergamon.

Gebhard, J. G. (1999). Seeing teaching differently through observation. In J. G. Gebhard & R. Operand (Eds.). *Language teaching awareness: A guide to exploring beliefs and practices* (pp. 35-58). Cambridge: Cambridge University Press.

Gebhard, J. G., Hashimoto, M., Joe, J., & Lee, H. (1999). Microteaching and self-observation: Experience in a preservice teacher education program. In J. G. Gebhard & R. Oprandy (Eds.). *Language teaching awareness: A guide to exploring beliefs and practices* (pp. 172-194). Cambridge: Cambridge University Press.

Gebhard, J. G., & Ueda-Motonaga, A. (1992). The power of observation: "Make a wish, make a dream, imagine all the possibilities!" In D. Nunan (Ed.). *Collaborative language learning and teaching* (pp. 179-191). Cambridge: Cambridge University Press.

Genesee, F., & Upshur, J. A. (1996), *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hitchcock, G., & Hughes, D. (1995). *Research and the teacher: A qualitative introduction to school-based research* (2nd ed.). New York, NY: Routledge.

Long, M. H. (1980). Inside the "Black Box": Methodological issues in classroom research on language learning. *Language Learning, 30(1)*, 1-42.

Lynch, B. (1996). *Language program evaluation*. Cambridge: Cambridge University Press.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Richards, J. C. (1998). Through other eyes: Revisiting classroom observation. In *Beyond training*. Cambridge: Cambridge University Press.

Rossi, P. H., Freeman, H. W., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Newbury Park, CA: Sage.

Spada, N. (1990). Observing classroom behaviours and learning outcomes in different second language programs. In J. C. Richards & D. Nunan (Eds.). *Second language teacher education*. Cambridge: Cambridge University Press.

Spada, N., & Lyster, R. (1997). Macroscopic and microscopic views of L2 classrooms. *TESOL Quarterly 31(4)*, 787-792.

Stern, H. H. (1989). Seeing the wood and the trees: Some thoughts on language teaching analysis. In R. K. Johnson (Ed.). *The second language curriculum*. Cambridge: Cambridge University Press.

Vierra, A., & Pollock, J. (1992). *Reading educational research*. Scottsdale, AZ: Gorsuch Scarisbrick.

Wang, Q., & Seth, N. (1998). Self-development through classroom observation: Changing perceptions in China. *English Language Teaching Journal, 52(3)*, 205-213.

Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford: Blackwell.

Williams, M. (1989). A developmental view of classroom observations. *English Language Teaching Journal 43(2)*, 85-91.

Woods, D. (1996). *Teacher cognition in language teaching*. Cambridge: Cambridge University Press.

CHAPTER TEN

DATA FROM DIARIES AND JOURNALS

In this chapter you will learn three ways of looking at diaries and journals, which topics diaries and journals have been used to investigate, and how you can use a diary or journal for your own research purposes.

While the first accounts of diary studies date from the 1970s (Schumann & Schumann, 1977), the use of written reflection on personal learning has only been popular in ESL classrooms from the 1980s onward (Bailey, 1991). Journals and diaries have been used to facilitate written dialogue between students and teachers, to assist self-assessment by teachers in training, and enable individual researchers to investigate their own learning processes. At the same time, the terminology to describe these instruments has grown increasingly confusing. Researchers variously refer to these instruments as diaries, journals, letters, logs, or some combination of these terms, and these instruments, by whatever name, have been used to investigate various topics such as evaluation (Parkinson & Howell-Richardson, 1989), learner characteristics (Bailey, 1995), learning context (Schumann & Schumann, 1977), learning processes (Schmidt & Frota, 1986), student self-assessment (Griffee, 1997), teacher reflection (Thornbury, 1991), and the writing process (Holmes & Moulton, 1995).

What is a general description or definition of a diary?

The most popular term in the literature is some form of the word diary—here is a list of its variants:

- *Diary studies* (Bailey, 1995; Campbell, 1996; Jones, 1994; Matsumoto, 1989, 1996; Schmidt & Frota, 1986)
- *Learner diary* (Lowe, 1987; Parkinson & Howell-Richardson, 1989)
- *Language learning diaries* (Peck, 1996; Schumann, 1980)
- *Teacher diaries* (McDonough, 1994)

The next most popular term is some form of the word journal:

- *Dialogue journal* (Carrell, 1990; Casanave, 1995; Holmes & Moulton, 1995; Meath-Lang, 1990)
- *Student journals* (Casanave, 1994)
- *Journal studies* (Numrich, 1996)
- *Working journal* (Spack & Sadow, 1983)
- *Journal* (Cummings, 1996; Bailey, 1980)
- *Personal journals* (Meath-Lang, 1990)
- *Classroom journals* (Lucas, 1990)

A third term is log:

- *Listening log* (Kemp, 2010)
- *Teacher research log* (Griffee, 1995)
- *Teaching practice logs* (Thornbury, 1991)

If we look at these studies from the point of view of who was writing to whom, certain patterns emerge.

For example, Bailey (1991) divides diary studies into two categories: in the first category, the diarists and the analyst are the same person, in the second, the diarists and the analyst are different people. Matsumoto (1989) calls the first group *introspective* and the second group *non-introspective*. Casanave (1995) prefers journals to diary studies, but continues the dichotomy. Her first category (corresponding to Bailey's diarists and analyst as the same person and Matsumoto's introspective writing), Casanave calls *learning logs*. Her second category (corresponding to Bailey's second category where the diarists and analyst are different and which Matsumoto calls non-introspective), Casanave calls *dialogue journals*.

I propose the global term *journal writing,* which would be composed of three categories: *dialog journals,* defined as student-to-teacher writing; *teacher journals,* defined as teachers (or teachers in training) writing to a senior teacher or to each other; and *diary journals* defined as someone writing to herself. Figure 1 shows these relationships. The advantage of this model is that it simplifies the terminology, maintains the use of common terms associated with this type of writing, and relates them in ways that are familiar to language teachers.

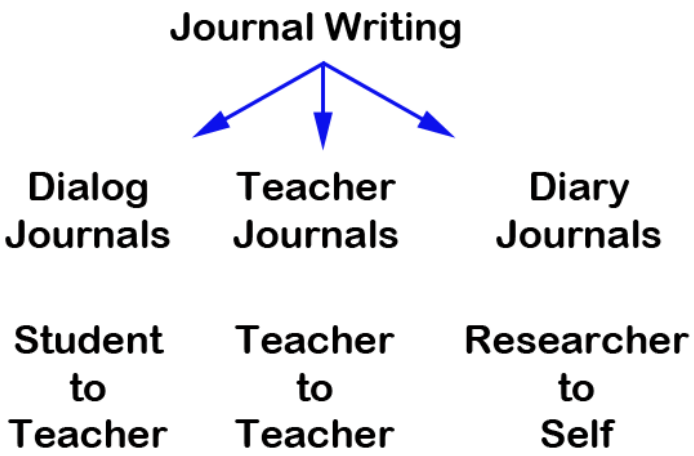


Figure 1. A model of journal writing showing three types

In this chapter, I will discuss journal writing using the three categories mentioned in Figure 1, namely dialog journals, teacher journals, and diary journals. For each type, I will discuss

advantages and disadvantages, key issues, and how you can do each type. I will also discuss data analysis, reliability, and validity for journal writing in general. Finally, I will conclude with a discussion of how diaries and journals can be used in research.

A word about the spelling of *dialogue* or *dialog*: *Webster's Student Dictionary* (1992) lists both spelling; the *Longman Dictionary of Contemporary English* (1995) lists *dialogue* as British English and *dialog* as American English. I will use the simpler American version, unless I am quoting another researcher who uses the more traditional spelling.

DIALOG JOURNALS (DJs)

Genesee and Upshur (1996) define a journal as a written conversation between students and teachers, while Ogane (1997) defines a dialog journal as a private written conversation between two persons. I would accept both definitions; however, the first definition applies to all journals and the second definition emphasizes the writing and personal aspect. In fact, a dialog journal could be audio recorded and DJs could be shared.

What form might a Dialog Journal take?

Many published reports of studies using dialog journals as data collection instruments don't report the physical aspect of the journal. Since Dialog Journals are typically a form of communication between one teacher and several students, convenience is important. For example, for Griffee (1997) and Ogane (1997), who were teaching large classes in Japan, the Dialog Journals were single sheets of paper turned in at the end of class. In a diary study I conducted in 2008, I passed out a piece of paper with short headings or questions followed by blank spaces for students to write in.

On advantages and disadvantages of journal writing

I will discuss advantages and disadvantages of each type of journal writing, but this discussion should not be interpreted as reasons for keeping or not keeping any of the types of journals. All data collection instruments have strong points (advantages) but also weak points (disadvantages). To use a traffic signal as a metaphor, an advantage can be interpreted as a green light saying go ahead, a disadvantage can be interpreted as a flashing yellow light saying be careful, slow down, think carefully about what you are doing because there are problems here, but then go ahead. There are no red lights in data collection, only green go-ahead lights and yellow caution lights.

What are the advantages of dialog journals?

Because Dialog Journals are between student and teacher, it can be argued that authentic communication can take place. It can also be argued that writing reduces stress (Ogane, 1997) and increases communication between student and teachers. Genesee and Upshur (1996) state that journals provide for increased communication, and thus provide feedback from students to teachers. Some students may feel reluctant to express their feelings about classroom learning publically in class, so journal writing may allow them to discuss these issues more privately. In this sense, journals provide an additional channel of communication between teacher and

student. Genesee and Upshur (1996) also argue that journal writing allows for the possibility that students may increase their interest in their learning through the reflection that journaling involves as well as allowing teachers to get to know their students better.

What are the disadvantages of dialog journals?

Ogane (1997) says there are no empirical data that writing in dialog journals causes improvement in reading and writing. This lack of empirical studies may be because to the extent that data from a diary or journal is introspective, it's hard to verify (Hilleson, 1996). Genesee and Upshur (1996) point out another possible disadvantage, namely that both students and teachers might need some orientation when first using dialog journals. In fact, some students may not participate in the journal process, or may participate against their will (Holmes & Moulton, 1995; Ogane, 1997).

What are some of the key issues to consider?

Since any type of journal or diary can be used for a variety of purposes (evaluation, pedagogy, increased reflection, self-assessment, or research), you have to decide why you might want to use a dialog journal. Following are some potential issues and the decisions they raise:

1. If you collect the dialog journals and use their contents to support your research, do you plan to correct or edit their contents in any way?
2. Do you plan to use the diary data with other forms of data, such as interviews and questionnaires? How do they mix and match? Does each form of data exhibit a unique feature or point of view, making it necessary? Can you articulate what each unique feature is?
3. In which language can the students write (Hilleson, 1996)? If they write in their L1, do you plan to translate it into another language for publication purposes?
4. Do you need to secure permission from your school to use data from DJs? If so, how should you go about doing that?
5. Do you know how will you identify various diary journalists? Will you use their first names, an initial or initials, or a pseudonym?
6. Will you assess the DJs in any way?
7. Does the teacher respond to the Dialog Journal, initiate, give information and opinions, or just ask questions (Ogane, 1997)?
8. When will Dialog Journals be collected, and who will collect them? For example, researchers Griffee (1997) and Parkinson and Howell-Richardson (1989) collected them at the end of each day.
9. Do you consider the Dialog Journal a private document between student and teacher or can DJs be shared? After each journal writing, Fedele (1996) told students they could share what they wrote. She had them (4^{th} graders) share with a partner. About halfway through the year, she corresponded with the students to keep them writing about a topic rather than abandoning it.

10. Genesee and Upshur (1996) raise the issue of regular times for writing entries, collecting journals, and reading them. Would writing take place at a regular time, and if so, what time?
11. How will you deal with ethical concerns if student express feelings such as sadness or depression? If students write something in their journals that indicates to you they may harm themselves or others, what would you do? Genesee and Upshur (1996) recommend knowing school policy on confidentiality and notification of authorities. If no such policy exists, decide on one of your own.

TEACHER JOURNALS (TJs)

A Teaching Journal requires not only a teacher but also students (otherwise it would be classified as a Diary Journal), and the students may be other teachers, graduate students, graduate-level or late-stage undergraduate native English speakers or non-native English speakers (otherwise it would be classified as a Dialog Journal). A typical example of what I am calling a Teacher Journal is produced when a teacher, often a university or college professor, engages students in a training project.

What form might a Teacher Journal take?

The physical form that a TJ might take is any convenient writing form, such as sheets of paper kept in a folder or a spiral bound notebook. Another way of thinking about the form of a TJ might take is its organizational form. For example, Francis (1995) had students keep journals in terms of four categories: teaching plans, reflective writing on the class, recording events that impacted their practice teaching, and a critical summary of their reading.

Thornbury (1991) gave his students seven categories for their diaries: the aim of the lesson, assumptions and anticipated problems, description of lesson, how the trainee felt, assessment of the effectiveness of the lesson relative to its aim, suggestions for improvement, and personal objectives for the next lesson.

What are the advantages of Teacher Journals?

One advantage of TJs is that they can help teachers investigate issues from a subjective or insider point of view (Bailey, 1990; Numrich, 1996). This means that TJ data can complement data from instruments, such as tests and questionnaires that investigate issues from more of an outsider point of view.

Another advantage is that TJs, like all diaries and journals, can engender reflection, which may in turn produce change. This may be because TJs can uncover processes by which novice teachers learn (Numrich, 1996). Teachers in training may even use TJs to ask questions, and then use the TJ to increase reflection, which may produce answers to those very questions (Porter, et al., 1990).

A third advantage is that TJs are low tech and easy to do. All that is required is the assignment,

something to write on, and perhaps some organizing questions or categories. A fourth advantage is that TJs can allow students to tell their instructor where they are having trouble because the TJ is a safe place to ask questions (Porter et al., 1990).

What are the disadvantages of Teacher Journals?

Bailey (1990) mentions that writing in a journal involves editing. This means that since we are not recording events as they happen, we are selectively remembering, or editing. Journal writers have to work hard to include objectively what happened as well as reflect on what they learned. For this reason, giving TJ writing categories as Francis (1995) and Thornbury (1991) did allows for the separation of what happened, what the journal writers felt about it, and what they learned from it.

What are some key issues for Teacher Journals?

Bailey (1990) mentions that confidentiality of participants must be taken into account. Increasingly, researchers are required by institutional review boards (IRBs) to submit their research plan for review and approval; one of the main concerns of IRBs is protection of research participants. Bailey (1990) also mentions participant resistance to keeping a diary. Researchers using Dialog Journals or Teacher Journals would be wise to have a policy in place for dealing with students who do not want to keep a journal.

DIARY JOURNALS (DJs)

You will recall that a diary journal is a document maintained by an individual writing a report to himself or herself on some topic area, such as learning a language or teaching a course.

What form might a Diary Journal take?

Some teachers keep a separate journal for each class in spiral-bound standard-sized notebooks (Isakson & Williams, 1996). It is also possible to keep a small note pad in pocket or purse for quick entries during the day, and then transfer a fuller account later. Others write on single sheets of paper that they number and keep in a folder, and of course diaries may be digital.

Why keep a Diary Journal?

There are two broad reasons for initiating a Diary Journal. One is a discovery or inductive reason, namely that we want to explore and learn something. For example, we are teaching a course, maybe for the first time, and we are wondering about the curriculum, our pedagogical practice, or our students' learning. We decide to keep a Diary Journal to record our thoughts and feelings that otherwise might be forgotten or overlooked when it comes time to evaluate and revise the course. On the other hand, we might have a specific problem we want to examine. This is more a deductive reason. For example, we want to investigate a particular problem or issue that has been on our mind, or that we have been reading about. Ellis (1989) used diaries in this way. He asked two teachers to keep diary journals to reflect on their learning. He then compared the results of the diary journals with the results of a questionnaire to investigate learning styles completed by the same two teachers.

What are the advantages of keeping a Diary Journal?

One advantage is that, as noted by Bailey (1995), being a diarist is somewhat like being a participant observer in an ethnographic study. If someone wants to be a participant observer, that is to say to be both participant in a situation and at the same time to be able to observe what happens, one has to either ask for permission to be there, or one has to somehow sneak into the situation, hopefully unnoticed. But as a teacher in a course, you are perfectly situated because you are already there. You already have access to one of the main participants who, in the case of a diary journal, is you.

Another advantage of a diary journal is that there is a growing consensus that writing creates reflection, and reflection in and of itself is helpful (Cooper, 1991; Holly, 1989; Hoover, 1994). Journaling of any kind can be a rich source of data for whatever use we are seeking (McDonough, 1994). This can be helpful in research, especially extended research such as working on a thesis or dissertation. Borg (2001) documents how he used a diary to examine his own research and writing problems, and by examining and becoming aware of certain tendencies he was able to change them.

What are the disadvantages of keeping a Diary Journal?

Diary Journal entries, because they are about the self, have the potential to be personal or even painful (Bailey, 1995). Of course, no one knows about the diary except the person keeping it, and as McDonough (1994) points out, the diary data must be analyzed. The uncomfortable aspect of the diary might arise at this point simply because we tend to be more accustomed to holding our students up for evaluation than we are holding up ourselves for critical analysis.

A second disadvantage of the diary journal is that, as Campbell (1996) notes, the original diary is not published. Usually only excerpts to support points are revealed. This makes comparisons and resulting conclusions difficult.

Finally, and more importantly, we are limited observers (Schmidt & Frota, 1986). We don't see or notice everything. In addition, as Fry (1988) says, diary data are retrospective data, in other words, we make diary entries after the event. That means there is some memory decay or lapse between the time the event occurred and the time we write about it. In addition, we have no way of knowing if we are typical observers. We know that we are influenced (biased) by our backgrounds, concerns, and interests. Would others notice the same things we did? It is for this reason that data from diaries of any kind must be considered data to be validated. It is not the case that diary data are valid simply by virtue of being written and recorded by us.

How can I begin my own Diary Journal?

Here are some possible suggestions to consider: First, just start. Find something with which to record your thoughts and begin. If a small audio recorder is handy and you don't mind listening to the sound of your own voice, press the 'record' button. If you want to use your computer, open a new document and start writing. If you prefer paper, find any notebook with some unused paper and begin writing. The essence of this tip is that starting something new is half the battle.

The downside of this approach is that "haste makes waste" and what you do not consider may turn out to be a problem. For example, if in your rush to begin a diary you forget to list categories that you might want to comment on, then later it may be difficult to include them. On the other hand, how will you know what the problems are if you don't begin? One way to deal with the "just do it" approach vs. the "think about it first" approach is to say to yourself, this is just a pilot. Do a pilot diary journal for the purpose of finding out how it works and what the problems might be.

On the other hand, if you are a "let's think about this" type of person, here are some things to consider. Are there any upcoming opportunities, such as beginning a new course or a trip you are about to take, that could be an opportunity for diary-keeping? What physical format such as type of notebook do you like? Are there any persistent problems that you have been wondering about that you could investigate with your diary? Where do you like to write? Do you have a special place to write, a special time, a special pen? Write out your questions and your answers to those questions. In doing so, you have just begun your diary.

What kind of data typically results from Diaries and Journals?

Diaries and Journals are virtual vacuum cleaners in that they can suck up anything. Data from journals can include daily events and related feelings (Schumann, 1980), impressions of people and circumstances in the environment of interest (Bailey, 1995), and critical inquiry focusing on practice (Holy, 1989). The data can take the form not only of words, but also of frequency counts, pictures, drawings, and even pieces of conversation overheard or transcribed from recordings.

How is Diary data typically analyzed?

The most common analysis advocated in the literature reviewed here is to read the journal entries and find patterns (Bailey, 1990, 1995; Borg, 2001; Hoover, 1994; Matsumoto, 1989; McDonough, 1994; Schumann 1980). Their advice is to read entries and notice themes, especially themes that reoccur. Then list these themes, perhaps with the number of times you notice them. This process, by the way, is similar to the statistical correlation procedures known as Factor Analysis, in which a factor would be the same thing as a theme. The only difference is that your mind, instead of a statistical procedure in a computer, is doing the sorting, but the results are similar: a list of identified themes, and in the case of journal data, a series of quotes to back up each theme.

A second approach to journal data analysis is illustrated by Schumann and Schumann (1977). They report that they asked one question: What did the data reveal that was not known before? They decided this question was not answerable. So they changed the question to: What did the data tell them about their language learning? This question was answerable, and it formed the basis for their results, which they labeled as themes, and later they called them personal variables. These variables stated the conditions under which each of them best learned a foreign language.

A third approach is to connect the data to theory. Although Bailey (1995) suggests rereading and looking for trends by frequency or saliency, she also says that data from journals can relate to theory without saying exactly how. She might be suggesting that one has a theory and can use data from journal entries to confirm or disconfirm the theory, or she might be suggesting that one can use data from journal entries to create a theory. Fry (1988) also allows for theory, in the form of a

framework, as useful for understanding data from journals. Fry (1988) raises three possible uses for data from journals: 1) to generate hypotheses, 2) to provide insights into learner strategies, and 3) to give insights into second language acquisition. He agrees that diary data can be used for the first use, generating research hypotheses. If the second use is desired, investigation of learner strategies, then he thinks is appropriate for the categories to emerge from the data. This is essentially the first approach to analysis. If the third use is desired, namely research into second language acquisition, then he wonders if journal data are appropriate. However, if that is the desired use, it will be necessary to describe the categories and framework carefully so other researchers can use them and verify them. Describing categories and framework amounts to describing the theory used. A fourth approach to analyzing data from diaries and journals is to analyze the data in terms of categories derived from discourse analysis. For example, Ogane (1997) used categories based on Peyton and Seyoum (1989), such as topic initiation and responses.

How can I calculate reliability and validity?

Reliability and validation evidence is seldom provided in diary study research. Perhaps such evidence tends to be ignored because subjective self-report data are assumed to be both reliable and valid. The reasoning is, if someone reports his or her own observations, thoughts, and ideas, how can that person be mistaken? I have listed many reasons in the disadvantages section of this chapter that suggests otherwise.

The main argument for investigating and reporting reliability and validity is that all data collection instruments have to account for it, and diaries and journals are not exempt. Because reliability and validity are seldom if ever mentioned in studies using data from diaries and journals, there is not much to draw from. What follows are some possible attempts that can be developed:

1. Triangulate diary data and recorded data of a similar type. Schmidt (Schmidt & Fronta, 1986) wondered how the occasional data he was gathering from his conversations in coffee shops would compare to more systematically collected data. He tape-recorded long stretches of conversation between him and his teacher (Fronta) so he could analyze and compare the two types of data. Another approach to validating journal data through triangulation would be to gather data from two different data collection instruments. For example, we could create a questionnaire that asked about learning styles, and also gather data from student journals on how they best learned in class. We could then compare interpretations from each of the instruments. If the conclusions we reached from one instrument matched or supported those of the other instrument, we could claim validation.
2. Stability of results over time can point to reliability. If the analysis of diary data consistently shows the same or similar results, these results could be an argument for reliability and thus validity of the data.
3. Bachman (1990) describes construct validity as the extent to which what we find in our data matches what we expect to find based on our theory. In other words, we can argue for construct validity of journal data if first, we define what we are looking for (our construct)

and second, show that our data and analysis are consistent with what we found. The key is to define the construct we are using.

4. Use peer review to establish validity (Miles & Huberman, 1994, p. 277). Peer review means finding a person with sufficient knowledge to understand the study; this person cannot be directly involved in the study, however. This person (the peer) can be asked to comment on their ability to trace a path from raw data to your conclusions. A series of three questions can be asked of a peer reviewer: Are my conclusions plausible? Can you think of alternative conclusions? Do you think my conclusions are not only plausible, but correct? These results can be offered as validity evidence.

How can diaries and journals be used in research?

Researchers conducting diary studies often wonder about validity, and their concern frequently focuses on the issue of generalizability (Bailey, 1980, 1991; Hilleson, 1996; McDonough, 1994; Matsumoto, 1989). One reason that generalizability is mentioned as lacking or impossible to obtain in diary studies, is because of the low N-size, that is, the number of persons involved in the study (Bailey, 1991). As a result, it is often lamented that the findings from diary studies do not, indeed cannot, generalize beyond the diary study itself. This is a misplaced argument because generalizability is a characteristic of a research design, not that of a data collection instrument, and a diary or journal is a data collection instrument, not a design. Understanding that a diary is a data collection instrument and not a research design relieves diarists from the burden of showing causality and generalizability of their data, but it increases the burden of showing reliability and validity of their interpretations based on that data. At the risk of being tedious, I will expand on this argument.

Generalization is the ability to apply the findings of a study or investigation to other situations that are reasonably similar to the original study. It is sometimes called external validity in contrast to internal validity, and is a concern of design—the blueprint of relationships within a research study. For example, survey research design shows the relationship of a sample to a population while an experimental design spells out the relationship of variables of interest in two (or more) groups or at two (or more) times within the same group. No matter what the design, there is a concern with showing how causality and generalizability are developed in research. A data collection instrument (DCI), on the other hand, is a way of collecting data. Examples of data collection instruments include questionnaires, interviews, and classroom tests. Unlike designs, which are concerned with causality and generalizability, DCIs are concerned with the quality of the data they collect, better known as reliability and validity. To use a metaphor of space, designs are like suns, some burning brightly, some in the process of dying, and some being born; DCIs are like planets, humbly revolving around suns.

Journals are data collection instruments, not designs

Keeping a diary is a form of collecting data, not a way of designing a research project, which would have to show causality and generalize results. As stated earlier in the Introduction to Research Design, any design can accommodate any type of data. A design without data is an empty shell, and data without a design is meaningless.

Data must be analyzed, as Bailey (1991) rightly points out, but data alone do not give a purpose or reason for collection, do not explain themselves (you need a theory for that), do not show causality, and do not generalize to other situations. Therefore, we can't compare data from diaries with the experimental research design because that would be to compare apples (a form of data collection) with oranges (a design).

How should diary studies be used? We have to recognize that diary data by alone are weak. This is not an attack on diary data as an inferior form of data; the same weakness can be said for any single form of data, including that from standardized tests or interviews. Given current research practices, it is doubtful that data from diaries and journals should be used alone. Probably, data from a journal are best used in a design for triangulation with data from other data collection instruments (see Ellis, 1989; Schmidt & Fronta, 1986), but this is not a point I would insist on.

Now we are ready to talk about the role of N-size in diary studies and the role of N-size in relation to generalizability. How can we take the insights found in the journal of one person and generalize those insights to the situation of others? The broader question is, how do data generalize, and what is the mechanism?

The authors of many diary studies accept the notion that samples generalize to populations, and samples suffer if they are few in number and are not typical of the population. For example, Bailey (1991) says that low N-size results in a situation that may inhibit generalization, and Matsumoto (1989) questions whether findings from diary studies may be generalized at all, since the data from a single informant may be idiosyncratic. This form of generalization, namely from a sample to population, is often attributed to experimental design, but in fact is that of survey design. Most designs, including the experimental design, use another form of generalization. They use data from their subjects, be it one or many, to create a *theory*. Statistics may or may not be involved, but if statistics are used, they are used internally in the study to show causality, not for purposes of generalizing to a population. In designs such as experimental and case study (but not survey design), the theory promotes or allows generalization to other situations, not the data from the number of persons involved. There is no reason why a diary study producing data from even one person cannot be used to create, confirm, or disconfirm a theory.

By way of summary and conclusion:

- A diary, journal, or log is a type of data collection instrument.
- Data resulting from diaries, journals, or logs are valuable forms of data, especially when used with other data collection instruments (tests, interviews, observation protocols, questionnaires) in the context of a design for the purpose of data triangulation.
- As with any data collection instrument, if used for research, data from journals require reliability and validation evidence.
- Results of research using data from diaries and journals can be used to create theory; from that theory, generalizations to other situations can be stated as hypotheses subject to further testing.

DISCUSSION QUESTIONS

Write some questions you had while reading about diaries and journals.

Task 1. Which of the three types of journal approaches discussed in this chapter (Dialog Journals, Teacher Journals, or Diary Journals) would be most interesting for you to try?

Task 2. Keep a journal on a topic of interest for one week. Make at least one entry per day on a topic of interest. At the end of the week, analyze your data.

Task 3. In the journal assignment in task 2, did you start from an inductive, bottom-up stance ("I just want to use a diary and see what happens") or from a deductive, top-down stance ("I know what I am looking for and I want to get some data on that topic.")?

References for Data from Diaries and Journals

Bailey, K. M. (1980). An introspective analysis of an individual's language learning experience. In R. Scarcella & S. Krashen (Eds.), *Research in second language acquisition* (pp. 58-65). Rowley, MA: Newbury House Publishers.

Bailey, K. M. (1991). Diary studies of classroom language learning: The doubting game and the believing game. In E. Sadtono (Ed.). *Language acquisition and the second/foreign language classroom* (pp.60-104). Singapore: SEAMEO Regional Language Centre.

Bailey, K. M. (1995). Competitiveness and anxiety in adult second-language learning: Looking at and through the diary studies. In H. D. Brown & S. Gonzo (Eds.), *Readings on second language acquisition* (pp. 163-205). Englewood Cliffs, NJ: Prentice Hall Regents.

Campbell, C. (1996). Socializing with the teachers and prior language learning experience: A diary study. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom* (pp. 201-223). Cambridge: Cambridge University Press.

Carrell, P. (1990). Reading in a foreign language: Research and pedagogy. *JALT Journal, 12*(1), 53-74.

Casanave, C. P. (1994). Language development in student journals. *Journal of second language writing, 3*(3), 177-201.

Casanave, C. P. (1995). Journal writing in college English classes in Japan: Shifting the focus from language to education. *JALT Journal, 17*(1), 95-111.

Cummings, M. C. (1996). Sardo revisited: Voices, faith, and multiple repeaters. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom* (pp. 224-235. Cambridge: Cambridge University Press.

Ellis, R. (1989). Classroom learning styles and their effect on second language acquisition: A study of two learners. *System, 17*(2), pp. 249-262.

Griffee, D. T. (1995). Student generated goals and objectives in a learner-centered classroom. *The Language Teacher, 19*(12), 14-17.

Griffee, D. T. (1997). Using dialogue journals for student self-assessment. In J. Johnson, E. Ogane & S. McKay (Eds.), *Working papers in Applied Linguistics 11*. Tokyo: Temple University Japan.

Holmes, V.L., & Moulton, M. R. (1995). A contrarian view of dialogue journals: The case of a reluctant participant. *Journal of Second Language Writing, 4*(3), 223-251.

Jones, F. (1994). The lone language learner: A diary study. *System, 22*(4), 441-454.

Kemp, J. (2010). The listening log: Motivating autonomous learning. *ELT Journal, 64*(4), 385-395.

Longman dictionary of contemporary English. (1995). Burnt Mill, Harlow: Longman.

Lowe, T. (1987). An experiment in role reversal: Teachers as language learners. *ELT Journal, 41*(2), 89-96.

Lucas, T. (1990). Personal journal writing as a classroom genre. In J. K. Peyton (Ed.), Students and teachers writing together: Perspectives on journal writing (pp 101-117). Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Matsumoto, K. (1987). Diary studies of second language acquisition. *JALT Journal, 9*(1), 17-34.

Matsumoto, K. (1989). An analysis of a Japanese ESL learner's diary: Factors involved in the L2 learning process. *JALT Journal, 11*(2), 167-192.

Matsumoto, K. (1996). Helping L2 learners reflect on classroom learning. *ELTJ Journal, 50*(2), 143-149.

McDonough, J. (1994). A teacher looks at teachers' diaries. *ELT Journal, 48*(1), 57-65.

Meath-Lang, B. (1990). The dialogue journal: Reconceiving curriculum and teaching. In J. K. Peyton (Ed.), Students and teachers writing together: Perspectives on journal writing (pp. 5-17). Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Numrich, C. (1996). On becoming a language teacher: Insights from diary studies. *TESOL Quarterly, 30*(1), 131-153.

Parkinson, B., & Howell-Richardson, C. (1989). Learner diaries. In C. Brumfit & R. Mitchell (Eds.), *Research in the language classroom* (pp. 128-140).

Peck, S. (1996). Language learning diaries as mirrors of students' cultural sensitivity. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom* (pp. 236-247). Cambridge: Cambridge University Press.

Schmidt, R. W., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237-326). Rowley, MA: Newbury House.

Schuman, F. M. (1980). Diary of a language learner: A further analysis. In. R. Scarcella & S. Krashen (Eds.), *Research in second language acquisition* (pp. 51-57). Rowley, MA: Newbury House.

Schumann, F. M., & Schumann, J. H. (1977). Diary of a language learner: An introspective study of second language learning. In H. D. Brown, R. H. Crymes, and C. A. Yorio (Eds.), *Teaching and learning English as a second language: Trends in research and practice* (pp. 241-249). Washington, D. C.: TESOL.

Spack, R., & Sadow, C. (1983). Student-teacher working journals in ESL freshman composition. *TESOL Quarterly, 17*(4), 575-593.

Thornbury, S. (1991). Watching the whites of their eyes: The use of teaching-practice logs. *ELT Journal, 45*(2), 140-146.

Webster's student dictionary. (1992). New York, NY: Barnes & Noble.