Luke Plonsky and Laura Gurzynski-Weiss

2 Research Methods

Abstract: This paper begins with the assumption that there is no perfect study. Rather, the process of conducting language acquisition research involves numerous decisions, each of which is accompanied by a set of strengths and weaknesses and which must be justified as appropriate to the substantive domain and the research questions being addressed. The chapter describes many of these choices and their corresponding benefits and drawbacks, illustrating key concepts and techniques with examples while making frequent reference to methodological issues and trends currently taking place in the field. In particular, we focus on major decisions related to (a) research designs, both descriptive/observational and (quasi-)experimental; (b) elicitation techniques and instruments for collecting data both linguistic (e.g., grammaticality judgment tasks) and non-linguistic (e.g., questionnaires) in nature; and (c) quantitative (e.g., analysis of variance) and qualitative (e.g., grounded theory) techniques for analyzing data.

Keywords: research design, quantitative research methods, qualitative research methods, language acquisition, statistics

1 Introduction

1.1 The Value of Methodological Knowledge

It is difficult to overstate the importance of methods in language acquisition research. Simply put, they are the means by which empirical evidence is gathered to advance knowledge. Our understanding of language acquisition is only as strong as the methods we employ. This path toward knowledge construction is contrasted with theoretical developments which complement, instruct, and are informed by empirical work (↗6 Language Acquisition Theories).

Beyond a complementary role with theory in advancing knowledge of language acquisition, the importance of understanding research methods can also be framed in relation to its value to different stakeholders in the field. Specifically, it is critical that both producers (i.e., researchers) as well as consumers (i.e., practitioners, policymakers) possess a keen understanding of research methods. With respect to the former, a conceptual and practical knowledge of how to conduct rigorous, principled, high-validity studies is key to accurately informing theory and practice.

A solid methodological understanding is also critical for non-researchers. Program directors, policy makers, test developers, and practicing teachers, among others, must possess a methodological foundation in order to appraise the validity of claims in the empirical literature and to make informed decisions. More specifically, consumers and others we might call “users” of empirical research must be able to attend to the many steps and decision points – along with their relative strengths and weaknesses – as presented by the author of each study. For example, recent inquiry into how language instructors seek out, interpret and use research findings in their language classrooms suggests that their use of this information is critical to the learning opportunities given to learners, and mediated by instructors’ individual characteristics including teaching experience and training (Borg 2013; Gurzynski-Weiss 2013).

In the broader field of applied linguistics, the importance of methodological knowledge has attracted increased attention in recent years, indicating momentum toward methodological and statistical reform (Plonsky 2014). The basic assumption underlying this movement is that methodological rigor and transparency are critical to advancing our knowledge of language learning and teaching. This broad notion as well as calls for greater attention to related but perhaps more nuanced issues (e.g., practical vs. statistical significance; the value of replication research) have been argued in meta-analyses and methodological reviews (e.g., Plonsky 2011; Plonsky/Gass 2011), surveys (e.g., Loewen et al. in press), journal guidelines (e.g., Chapelle/Duff 2003), and papers introducing novel techniques (e.g., Larson-Hall/Herrington 2010). The focus of this chapter is of course too broad to present a thorough discussion of the many individual issues that make up this movement. However, this paper would be incomplete without recognizing the momentum accumulating in this area. As we describe the methods of language acquisition research, we will highlight these and other positive changes taking place in the field.

1.2 Decision-Making and Trade-Offs

A starting point for understanding research methods is the fact that the process of designing and conducting a study is largely one of decision-making. For example, researchers must articulate the research questions they will address, the population they are interested in studying, the types of data they hope to obtain, and an appropriate set of analyses and interpretations.

The variety of methodological choices to be made is compounded by the vast number of substantive interests under the domain of “language acquisition”, an inherently interdisciplinary field with historical, theoretical, methodological, and practice-oriented ties to education, linguistics, and psychology, among other disciplines (↗1 Disciplines Relating to Language Acquisition). Having recognized these wide-ranging connections, it is worth bearing in mind that certain domains and types of research questions lend themselves more readily to certain methodological approaches. This fit between substantive interests and the methods employed to address them is what we refer to as “validity,” and will be seen throughout the chapter (cf. Mackey/Gass 2005, 106–119).

Finally, and closely related to this notion of validity, we want to stress early on that it is not productive or accurate to evaluate the choices made by a researcher as categorically "valid" or "not valid," "good" or "bad." Rather, each decision carries a unique set of strengths or weaknesses that must be weighed against the research questions, practical constraints, and the particular area in question. A classic example of this tension is the trade-off in lab- versus classroom-based research: Whereas research conducted in lab contexts may benefit from the ability to exercise greater experimental control over variables that might contaminate an experiment, these studies often lack ecological validity when compared to those conducted with intact groups in classroom settings.

2 Design Choices

While there are myriad ways to conceptualize methodological design, some – such as organization via type of knowledge being studied (e.g., more or less explicit; receptive or productive; declarative or proceduralized) – are more theory-specific and potentially variable in interpretation than others. For the sake of clarity, we present design choices as falling into one of two broad categories: (a) observation/descriptive studies; and (b) (quasi-)experimental designs. Each will be examined in turn alongside examples demonstrating how they have been used in language acquisition research. It is important to mention that study designs can be cross-sectional, meaning that data are collected at one point in time, or longitudinal, meaning that data are collected over a period of time.

Two critical decisions when designing a linguistic study are: (a) what will the research question be? and (b) how can I most robustly answer this question? (cf. Mackey/Gass 2005, 16–21). The methods and instruments used to answer a research question determine a study's validity and, hence, its contribution to our understanding of language acquisition.

2.1 Observational/Descriptive Designs

Studies that are observational/descriptive in nature seek to describe a linguistic phenomenon or set of phenomena without attempts at incurring change. These studies do not test the effects or outcomes of a treatment or intervention; they simply look to describe as systematically as possible what is occurring, interpret the occurrence within a larger context and, at times, hypothesize why the incidence is observed.

Two areas of acquisition research that independently and frequently utilize these designs, amongst others, are Hispanic sociolinguistics and individual differences. In the former area, for example, researchers examine which features predict code switching within a speech community (e.g., Poplack 2001), or how children learn sociophonetic variation (e.g., Díaz-Campos 2011). Individual difference research, on the other hand, examines traits of learners, such as their working memory capacity (e.g., Mackey et al. 2010), or level of integrative motivation (e.g., Hernández 2006), and how these traits mediate language acquisition processes, such as processing novel phonological input or a student's desire to seek out learning opportunities outside of the classroom.

2.1.1 Qualitative Research Questions and Designs

Observational/descriptive studies are often qualitative in nature. Generally speaking, qualitative studies are more open and inductive than their quantitative counterparts, and often approach data holistically with open or loosely defined research questions (cf. Friedman 2011). Typically, researchers look for themes that emerge from the data and interpret them within the context in which they occur; most qualitative studies do not attempt to generalize results to other populations. The goal in qualitative research is to have sufficient data to describe a phenomenon in detail, and this often occurs via triangulation (the use of multiple data sources) to ensure comprehensive and robust data collection.

An example of a qualitative descriptive study is Gurzynski-Weiss (in press), which examined how graduate instructors of Spanish make moment-to-moment feedback decisions. Data from video and audio recordings, stimulated recalls, and questionnaires were triangulated to develop a thorough understanding of instructor cognition. Many qualitative studies begin with broad research questions and, via multiple iterations of data coding and analyses (discussed below), edit their research questions based on the trends that emerge from the data. For example, Gurzynski-Weiss started with a general research question, "How do Spanish instructors make feedback decisions during non-experimental lessons?" and found that instructors continuously cited factors that influenced their decisions. These factors, based on themes emergent from the data, turned out to be both internal and external to the learners in the class, and the research questions were subsequently fine-tuned to reflect this.

2.1.2 Quantitative Research Questions and Designs

Other observational designs aim to collect and describe quantitative data, centering on how much or how often a phenomenon occurs. In this type of research, the focus is still on description and contextualization of data, but related to frequency of occurrences.

Ellis/Basturkmen/Loewen (2001) is a clear example of a descriptive, quantitative design. The researchers examined the quantity of naturally occurring focus on form (grammar within meaning-based context), and the breakdown of how many instances were teacher- versus learner-initiated. Thus, this study is both descriptive in the sense that it does not try to influence the linguistic environment, and quantitative, in that it focuses on the amount and initiator of focus on form episodes.

Although data in observational/descriptive research may be collected at more than one time, researchers do not compare results before and after a treatment. When researchers are interested in examining the effects of how a treatment might instill change, they turn to experimental designs.

2.2 Experimental Designs

A second broad category of designs involves the provision of a treatment, the effects of which are measured comparing pre- and posttests, and which often have relevance to language teaching. In doing so, experimental research addresses questions central to both theory and practice, such as "How effective is explicit grammar instruction compared to implicit when learning the subjunctive?" (cf. Loewen/Philp 2012).

Regardless of the particular questions being asked in an experimental study, there are choices to be made regarding the design. There is no perfect study, and each of these features or components may improve or detract from the findings. For the sake of brevity, three major features are described: pretesting, control/comparison groups, and delayed posttesting.

2.2.1 Pretesting

The main advantages for pretesting are two-fold: First, a pretest enables the researcher to measure and compare the participants' knowledge or behavior before and after treatment. And second, when a control or comparison group is included in the design, a pretest can be used to demonstrate group equivalence prior to the treatment. Without a pretest, it is difficult to know whether an advantage for one group can be reliably attributed to a treatment condition or to differences present prior to the intervention.

In certain situations, a pretest may not be necessary or even preferable such as when all participants are assumed to be uniformly naïve to the target structure. For example, Ellis/Sagarra (2011) were interested in learners' responses to different training conditions in Latin. Because the researchers excluded participants who had knowledge of the target language, no pretest was given. In other situations, researchers may choose not to pretest in order to avoid alerting the participants to the target feature or to avoid practice effects that may result from multiple administrations of the instrument. A third reason researchers may choose not to include a pretest in an experimental study is because they have assigned participants randomly to treatment conditions, enabling researchers to assume comparability across groups and link posttest differences to treatment effects. Despite the benefits of this design feature, random assignment at the individual level can be difficult if not impossible in research that is classroom-based and/or relies on intact samples.

2.2.2 Control/Comparison Groups

Researchers often choose to include a control or comparison group to show that the change or learning that has taken place is due to the treatment itself, isolated from other influences. Plonsky/Mills (2006) demonstrated that their treatment led to a significant change in perceptions of written feedback among college learners of Spanish. However, because there was no control group, it is unclear whether or not the change was due to the treatment. Variables such as the style of feedback provided may have also influenced students' perceptions over time.

The labels "control" and "comparison" are both used to refer to a group of participants who do not undergo the experimental treatment and whose data can therefore be compared to the treatment group. The choice of one or another, however, represents an important distinction and yet another critical design choice to be made in experimental studies. Whereas a comparison group generally receives a traditional or minimal intervention, a (true) control receives no treatment whatsoever.

Choosing which condition(s) to include involves a solid understanding of the substantive domain under investigation. Early research in a particular area may warrant the use of a true control group as a means to determine whether further inquiry is warranted. Once the effectiveness of a particular type of treatment is more established, comparison groups along with the use of multiple treatment groups can be used to test variations of different conditions vis-à-vis a more typical or traditional treatment. For instance, in an attempt to isolate the effects of explicit information (EI) in processing instruction, Fernández (2008) included and compared two treatment groups: one that received the typical processing instructional with EI and another that only received structured input and no EI.

In most domains, the domain is neither completely developed nor in an incipient state. In these cases, if the logistics allow, both conditions can be included. Such a design enables the researcher to estimate the effect of the treatment both as an absolute value as well as in relation to another treatment (i.e., as compared to the comparison group).

2.2.3 Delayed Posttests

In addition to measuring treatment effects against a control condition, experimental researchers are often interested in the longevity of those effects as well. In such cases, a delayed posttest is used. The findings provided by a delayed posttest can have important implications in applied contexts by informing the pace or interval at which new material should be introduced or recycled. Delayed effects can also be weighed against the resources needed to induce them (e.g., time, experimental manipulation) as a measure of their practical significance (Plonsky/Oswald forthcoming).

In addition to these benefits, the choice to include a delayed posttest carries with it several choices and potential drawbacks. For example, the researcher must determine an appropriate interval for administering posttests, considering previous research, the predictions of theory, and practical constraints, among other factors. Second, as with pretests, the internal validity of delayed posttests may be compromised by practice effects. A third potential threat to internal validity involves participant attrition. The longer the study, the greater likelihood that participants will be absent or fail to complete it, thus creating a potential for bias because better performing participants are less likely to attrite. Finally, including one or more delayed posttest requires additional time. If the research takes place with intact groups, the benefit of obtaining results for delayed effects must be weighed against the loss of valuable class time.

In our opinion, despite these risks, the balance generally falls in favor of the potential of delayed posttests to inform theory and practice. Nevertheless, we find this feature in a relatively small portion of experimental studies on language acquisition: only 38% in Plonsky's (2013) review.

3 Data Elicitation Choices

In this next section we present some of the main techniques, tools, and instruments commonly used, highlighting the pros and cons of each. Techniques whose primary focus is linguistic are discussed first, followed by those considered to be non-linguistic, though the categories are not always mutually exclusive.

3.1 Linguistic Focus

Instruments with a linguistic focus have either been designed or manipulated to elicit a speaker or learner's receptive or productive knowledge of a language. Receptive knowledge targets one's ability to recognize or choose the correct form from a list of options; it is believed to come prior to productive abilities in a second language. Productive knowledge refers to one's ability to produce the target language.

3.1.1 Grammaticality Judgment Tests

Utilized by acquisition researchers of almost all theoretical backgrounds, grammaticality judgment tests (GJTs) are a metalinguistic judgment task that requires a participant to judge the grammaticality – and often the acceptability – of an isolated sentence or set of sentences. A grammatical statement adheres to the rules of a particular language or dialect. An acceptable statement or question, on the other hand, violates one or more grammatical rules but is utilized in common speech.

Within the GJT instrument, methodological choices abound. GJTs are often presented via a computer visually, aurally, or in combination. Speakers judging the grammaticality of a sentence may simply be asked to press a computer button when they hear or read something ungrammatical, or they may be asked to press a certain button for a grammatical sentence, and another for an ungrammatical sentence; participants can also be asked to correct any sentences they find ungrammatical. Those asked to provide acceptability judgments can likewise be asked to differentiate between acceptable and unacceptable sentences, or may be asked to rank the acceptability of a sentence on a scale. GJTs can also be presented with or without time pressure. These seemingly subtle variations of this instrument importantly differentiate the specific constructs being tested. Timed GJTs test a participant's implicit or proceduralized knowledge of a language, as the participants are theoretically unlikely to be able to explicitly process rules, instead relying on their grammatical intuition. Untimed GJTs, on the other hand, tap into more explicit or declarative knowledge, giving participants time for thinking metalinguistically.

GJTs allow a high level of control for the researcher, and ensure that all participants are tested on the same grammatical items. However, the sentences are presented in isolation of contextual information, and are often artificial or infrequent in everyday speech. Computerized GJTs are the most convenient and provide the most control over the data collection. When computers are not available, or researchers are investigating a language that is not written or asking for the participation of illiterate speakers, oral administration of GJT is used. In these cases additional care must be taken to keep the procedure as similar as possible between participants.

3.1.2 Free Constructed vs. Constrained Constructed vs. Selected Response Tasks

A second frequently utilized class of instruments are known as free constructed response, constrained constructed response, and selected response tasks. They share a common goal of eliciting production of linguistic data, either in oral or written form. The manner in which speech is elicited, however, is quite different. Free constructed responses are most often communicative tasks, where participants are free to construct their own response to a prompt. For example, directing their classmate from point A to B on a map is considered to be a free constructed response task. Free constructed response tasks encourage authentic language production. That is, participants are free to utilize all linguistic tools in their repertoire, focusing on the meaning of their communicative message above grammatical form, echoing how language is used outside of experiments. However, because language production can vary considerably, there is no guarantee that the linguistic target item will be produced, and the issue of how to rate such open speech is challenging (Erlam 2006). Thus, when designing instruments researchers must ensure that the linguistic items are essential to successful completion of the task.

Constrained constructed responses also have a focus on participant production of linguistic output, but there is less freedom in the possible response. Participants completing this type of task are asked to fill-in-the gap(s) within a sentence (or series of sentences) with vocabulary and grammar appropriate to the surrounding context, predetermined by the researcher. For example, participants may be asked to complete a paragraph about the series of events a character did this past weekend. Because the events all happened in the past to a third person, the participant is provided cues within the text to construct their responses within the confines of a particular tense and with respect to the third person. The overall focus is still on meaning, but accurate production plays a greater role than in free constructed response.

Finally, selected response provides the most controlled option in this set of instruments. These tasks ask participants to choose from a provided set of options, either within or immediately following the gap (e.g., "circle one"), or they may be given a list of options in the style of multiple choice. Giving participants a choice tests their receptive knowledge of the linguistic target, as they are simply asked to determine which response fits better in the space rather than produce evidence of their knowledge. An advantage of selected response tasks is that they offer the most control over participant response and are easy to enter into common statistical software, such as SPSS. A disadvantage to this technique is the risk that participants are not interpreting the question and/or options in the same manner, and thus their responses, while superficially congruous, could in fact be dissimilar. At the same time, an undetermined percentage of participant responses may be correct simply due to chance, thus providing an inflated view of their knowledge. For these and other reasons, pilot testing is exceptionally important for this technique (and recommended for all others, as well).

3.1.3 Elicited Imitation

Elicited imitation tasks require participants to repeat a sentence, most commonly in the oral mode. The idea behind this task is that the inaccuracies in the participant's production will reflect target features that have yet to be fully acquired. Uses of this instrument include measuring native and L2 proficiency, among others (cf. Tracy-Ventura et al. 2014).

One key component of an elicited imitation task is the length and structure of the cue sentence (the sentence to be repeated), as reconstruction is considered to be a necessary component for a true measure of proficiency, rather than rote repetition (cf. Erlam 2006; Vinther 2002). Another component that can be altered is the timing of the repetition: immediate or delayed. Last, the main criticisms of this tool are trifold: the potential for imitation rather than reconstruction, the artificiality of the task, and the risk and variation of possible interpretations of learner errors.

3.1.4 Discourse Completion Task

Discourse completion tasks are used widely in sociolinguistics and pragmatics research for both first and second language acquisition. These tasks ask a speaker or learner to respond to specific prompts, either in written, oral, multiple choice, or fill-in-the-blank form (cf. Pavaresh/Tavakoli 2009). Discourse completion tasks are ideal when a researcher wants to examine a response to a specific type of speech act, or have a speaker/learner respond to a specific prompt. Due to the control offered, the researcher can manipulate various factors, such as the age, power, and gender within the discourse to which the participant is responding. This allows the researcher to measure a learner's acquisition of L2 pragmatics and sociolinguistic variety, and provides the researcher with a valuable tool for collecting native speaker data as a baseline for comparison with learner data. Discourse completion tasks are also useful for examining dialectal and sociolectal differences within a speech community. Main criticisms of these tasks are the potential artificiality and isolation in the presentation of discourse (e.g., Golato 2003). Researchers must be careful to make the scenarios as authentic as possible and take into account potential ambiguities which may compromise results.

3.1.5 Corpora

A corpus of linguistic data serves to provide a common pool of data for a variety of experiments and studies. The amount of data is usually considerable, although it varies from one corpus to another, and is usually stored electronically. The subjects of corpora can be highly specific, such as Chilean middle class radio speech, or quite broad, such as Davies' (2012) Spanish corpus (http://www.corpusdelespanol.org), which houses over 20,000 Spanish texts (100 million words) from an 800-year period. The strengths of corpus data are both in the quantity of data, as well as in the shared nature of the database. So often in language acquisition research we are limited to small-scale studies wherein target features occur relatively few times. Corpus data, on the other hand, often allow for tracking linguistic development on a much larger scale, and permit researchers from various backgrounds to investigate the same data pool from multiple perspectives (e.g., Asención-Delaney/Collentine 2011).

3.1.6 Ethnography and Diary/Journal Studies

Ethnographies are longitudinal observational studies whose aim is to acquire a rich detailed account of language use within a particular community (cf. reviews in Harklau 2005). Ethnographic researchers typically triangulate data from multiple sources, such as observations, interviews, questionnaires or surveys, and video and audio recordings.

Diary/journal studies typically ask one language learner or small group of learners to journal over a certain period of time. These studies can include prompts, asking learners to write about a specific theme, have a pre-determined quantity and frequency of required production (e.g., write one page every day), or can be completely open. Diary/journal studies are also often designed in conjunction with additional data elicitation methods such as informal or semi-structured interviews, questionnaires, or pre- and post-experience measures of proficiency. This is especially useful when examining a holistic experience, such as study abroad or when attempting to identify factors that make a certain context more beneficial for learning. For example, Schmidt/Frota's (1986) seminal study, which inspired the Noticing Hypothesis (Schmidt 1990), employed journal entries to describe the acquisition process. By journaling his learning of Portuguese, Schmidt found his noticing of linguistic structures was what ultimately determined which aspects of the language were acquired successfully, and which were not.

3.2 Non-Linguistic Focus

Instruments considered to have a "non-linguistic" focus are those that are used in many social sciences and are not specific to the study of language and language acquisition.

3.2.1 Questionnaires and Surveys

Questionnaires and surveys are common elicitation instruments in many domains (cf. Dörnyei/Taguchi ^{2}2009). Although some are still administered in-person, many are being transferred to online formats. Surveys are relatively easy to create and offer myriad types of questions such as multiple choice, ranking, Likert-scale, open ended, and essay questions. Surveys, particularly those administered online facilitate efficient coding, immediate visualization of trends, and can be entered directly into software designed for qualitative and/or quantitative analysis. Questionnaires can be disseminated online via email, Facebook, and list-serves. While this ability increases the potential generalizability of the results, this distance also increases the risk that participants interpret questions differently, particularly if the instrument was piloted in a culture different than that of the target population. For this reason, pilot testing with a diverse participant pool is vital to ensure robust data collection.

3.2.2 Measuring Reaction or Reading Time

Reaction and reading time data can provide information regarding how the learner is processing information. For example, as discussed with GJTs, faster reaction times can provide an indication that the learner is tapping into implicit knowledge or has automatized knowledge. Slower reaction and reading times provide an indication that the particular linguistic structure(s) are not fully learned. How to determine what constitutes a "fast" or "slow" reaction or reading time, however, is challenging. There are individual learner or speaker differences; some people simply read faster or slower than others and there are potential differences across languages as well (cf. Jiang 2012).

3.2.3 Observations and Interviews

Observations are usually characterized as open/unstructured or closed/structured. In open observations, the researcher is present in an acquisition environment, such as a classroom. The researcher often takes notes, makes audio and/or video recordings, and collects supporting information (such as a lesson plan and homework) that focuses on a specific linguistic target (e.g., the future simple), several targets (e.g., preterit versus imperfect), or on the general interactions within the space. A structured observation, in contrast, has a pre-determined scheme for collecting data, such as via a grid or chart. The COLT, Communicative Orientation to Language Teaching (Spada/Frölich 1995) is one such option for a structured observation. The COLT contains spaces alongside key factors in classroom interaction, such as initiation of discourse and use of the target language, and the timing of each task. Open observations allow for themes to emerge naturally, while structured observation techniques allow for comparability between datasets and provide a scheme that helps the researcher stay focused on the task at hand. With observations, it is important that the researcher's presence is not an influence (i.e., the "observer's paradox"; Labov 1972).

Interviews, on the other hand, can be categorized as unstructured, semi-structured, or structured. Unstructured/open interviews are akin to a conversation between researcher and participant. Here the interviewer does not have a specific set of questions or script to follow, but poses general questions to start the conversation. For example, in a study about language teaching experience, a broad question such as "Tell me about your teaching" may initiate the discussion. During the interview, topics of interest are explored under the guidance of the researcher. For this reason the researcher must be experienced or trained in interacting with the participant population and be clear on the research questions at hand. In semi-structured interviews the researcher may have a few specific questions. S/he may spend more time on one topic than another and may explore allied topics that arise as well. The researcher must have adequate experience to know which topics should be explored further for the project at hand, and how to bring the interview back to the main topic.

Structured interviews offer the most control and follow a pre-determined set of questions or script that is asked of each and every participant. As discussed previously, the more control the researcher has, the more guaranteed he will get the data needed to answer the research questions. Whichever technique is utilized, in order to encourage reliable, authentic data, it is paramount that the participant is comfortable with the researcher and the questions asked. They must be posed in such a way that the participant feels free to answer honestly, rather than answer in an attempt to please the researcher or match the study's interest. Having some flexibility in the interview schemata and starting with less-personal questions can assist in ensuring this comfort. While space prohibits their discussion in detail here, for an additional retrospective measure the reader is directed to stimulated recall protocols (cf. Gass/Mackey 2000; for concurrent introspective measures on participant within-task cognition, think-aloud protocols cf. Bowles 2010).

4 Analyzing Data

Once data have been collected, researchers move on to the analysis phase. As in related disciplines, language acquisition researchers depend on wide variety of analytical approaches, most of which are well established and shared across the social sciences. These techniques are often categorized as quantitative or qualitative, although many blend the two (i.e., "mixed methods"), a decision that can yield a rich depiction of the phenomena in question (cf. Hashemi/Babaii 2013). We preserve this distinction here, nevertheless, for the sake of simplicity.

It should also be mentioned that we have necessarily omitted numerous concepts and techniques. In the quantitative domain, we have focused on descriptive statistics, which are often overlooked, and on analyses comparing means, which are exceedingly common in the field. On the qualitative side, we focus on grounded theory (Glaser/Strauss 1967), discourse analysis (Foucault 1981), and computer-assisted qualitative data analysis software. For more specialized treatments of these topics, we recommend readers consult Mackey/Gass (2012), and Richards/Ross/Seedhouse (2012).

4.1 Quantitative Analyses

Prior to conducting statistical analyses, researchers must calculate and examine their descriptive statistics. These include means, standard deviations (*SD*s), confidence intervals, and frequency counts. In many cases, tests of statistical significance contribute little or nothing beyond what can be revealed in very basic summary or descriptive statistics, especially when combined with visual inspection of the data (e.g., box plots, scatter plots; Larson-Hall/Herrington 2010; Larson-Hall/Plonsky forthcoming). Imagine, for example, an experimental study comparing the effects of a traditional and novel treatment wherein the average of the experimental group's posttest is 34 (*SD* = 4) and the comparison group 26 (*SD* = 4). It is clear without formal testing that the novel treatment is more effective.

Another descriptive statistic that has received considerable attention in recent years is the "effect size", which indicates the strength of an intervention or of a relationship between variables (Plonsky 2012). Cohen's *d*, one such effect size index, expresses the mean difference between groups in *SD* units. (Other common indices include *r, r*-squared and eta-squared.) Returning to the previous example, we could quantify the difference in effectiveness between the two treatments by a *d* of 2.00. This result, generally considered quite large in the realm of L2 research, confirms what we were able to tell from looking at the groups' means and *SD*s. Another benefit of effect sizes such as *d* is that they are standardized and therefore can be compared and combined across studies via meta-analysis (Norris/Ortega 2006; Plonsky/Oswald forthcoming). Finally, effect sizes can also be used to inform the design of future studies by enabling a priori power analyses (Cohen ^{2}1988).

Although the importance of descriptives cannot be overemphasized, there are times when tests of statistical significance are useful in identifying patterns in the data. Specifically, such analyses help us determine whether the results are indicative of a real relationship between variables or perhaps spurious (i.e., due to chance or other factors such as instrument reliability). Statistical tests can be especially useful with multivariate datasets, when interactions between variables may be present. The remainder of this section focuses on statistical analyses for comparing means, the most common type in the field (Plonsky 2013).

Although there are several analytical options for comparing means across one or more groups or independent variables, the choice of test is fairly straightforward. Assuming the data conform to a normal distribution (yet another reason to examine the descriptives), *t* tests are used to compare means between two groups, and analysis of variance (ANOVA) is used for three or more groups. Of course this is an oversimplification. There are several varieties of *t* tests and ANOVAs, such as independent samples (i.e., between groups) and dependent samples (i.e., one-sample / pre-post). Another variant of the ANOVA, the analysis of covariance (ANCOVA), enables the researcher to control or adjust for a covariate such as a pretest score or a measure of aptitude. Still another, multivariate analysis of variance (MANOVA), allows for simultaneous comparison of means across multiple dependent variables. Ortega (1999), for instance, used MANOVA to compare multiple measures of interlanguage performance (e.g., words per utterance, type-token ratio) of participants in two conditions (+/– planning).

Because of the predominance of analyses comparing means, we feel the need to express our concern, if only briefly, over three problems associated with this analytical approach: (a) incomplete reporting of data (often *SD*s to accompany means); (b) unmet and unchecked assumptions (e.g., a normal distribution); and (c) the tendency to choose statistical tests based on convention or convenience rather than for their appropriateness to the data and questions being posed (Plonsky 2013; Plonsky/Gass 2011).

4.2 Qualitative Data

In addition to a thorough examination of the data after collection is complete, analysis in qualitative research usually also occurs throughout the collection process. Prior to the primary analyses, qualitative data are often transcribed using transcription conventions standard to the theory in which the study is based. These transcriptions are then coded thematically based on the research questions at hand and the theoretical orientation of study. For this reason robust operationalizations and coding protocols are paramount to predicating vigorous qualitative research designs, even more so than in quantitative studies, as there is often more room for interpretation than with studies centralized on numbers. For the sake of space, three types of qualitative data analysis will be discussed: grounded theory/content analysis, discourse analysis, and computer-assisted qualitative data analysis software. Many additional analyses are available to researchers and, as the field of language acquisition is rapidly increasing its use of qualitative and mixed methods, so too are the options for analysis.

4.2.1 Grounded Theory/Content Analysis

In grounded theory (Glaser/Strauss 1967), also known generally as content analysis, the coding protocol develops ground-up from within the content of the data rather than from an imposed top-down approach, and occurs via several iterations of coding. Initial coding is open, and the researcher examines and begins to code/categorize the data. In the next rounds the researcher codes across and within participants for recurrent themes. Finally, the researcher settles on a coding scheme, develops a coding protocol, and revisits the data with operationalizations that have typically emerged from the dataset. This type of analysis is highly specific and by design the appropriateness to the particular dataset is guaranteed. Additionally, if operationalized clearly, results can be discussed in relation to other studies that have examined similar themes and followed a similar grounded theory analysis.

4.2.2 Discourse Analysis

A second frequent type of analysis of qualitative data is referred to as discourse analysis (Foucault 1981), which is a broad term encompassing the study of linguistic or communicative structure of speech rather than being limited to thematic content. A researcher may examine, for example, how learners of different proficiency levels take turns during dyadic interaction, how power is established within a given speech community, or which instances of *ser* beginning learners of Spanish have mastered and which are still being confused with *estar*. As with grounded theory, transcriptions tend to be a starting point for studies analyzing discourse. During initial coding, researchers focus on contextual factors such as participants, setting, power relationships between participants, goals, and so on. The second waves of coding tend to center on linguistic and communicative structure of the interactive discourse (e.g., turn-taking, refusals). Finally, the researcher zooms in on specific patterns and features to form the analysis. An example from a study of Spanish pragmatics, for example, would be an examination of how learners who are native speakers of English negotiate an invitation refusal from a hypothetical boss (Félix-Brasdefer 2008). In this speech act the learner is faced with an interaction where the target culture is different from the native culture in both what is presented (multiple, seemingly insistent invitations) and what is expected (several polite emphatic refusals about how disappointed the student is but genuinely unable to attend).

4.2.3 CAQDAS

Although a full review is outside the scope of this paper, we felt the need to mention that researchers utilizing many different qualitative techniques have turned increasingly to computer-assisted qualitative data analysis software (CAQDAS). CAQDAS software allow for data management, easy and reliable searching of terms, and assistance with coding. For a review of CAQDAS programs and a step-by-step guide to one of the most widely used programs for second language acquisition research, NVivo, cf. Baralt (2011).

5 Conclusion

We have tried to emphasize throughout this chapter that conducting empirical research is, at its core, a process that involves myriad decisions. Therefore, whether taking the perspective of the individual researcher or the consumer of research, we must also judge language acquisition designs not as “good” or “bad” but, rather, in light of the strengths and weakness inherent in each technique and in the execution determined by the researcher’s theoretical orientation.

To be clear, we are not suggesting a type of methodological relativism. There are certain practices or design features that yield stronger evidence than other practices or features. Meta-analyses and methodological syntheses such as those cited in this chapter have brought many of these features and related issues to light such as the importance of pretesting, the value of effect sizes, and transparency in data reporting practices. The interest in methods brought about in these issues gives us hope for continued reform and increases in the field's methodological knowledge. Nevertheless, great strides are still needed in training and field-wide standards (cf. Plonsky 2014), among other areas, for language acquisition to move toward and to maximize its informational potential to theory and practice.

6 Bibliography

Asención-Delaney, Yuly/Collentine, Joseph G. (2011), *A Multidimensional Analysis of a Written L2 Spanish Corpus*, Applied Linguistics 32/3, 299–322.

Baralt, Melissa (2011), *Coding Qualitative Data*, in: Alison Mackey/Susan M. Gass (edd.), *Research Methods in Second Language Acquisition: A Practical Guide*, Malden, MA Wiley-Blackwell, 222–244.

Borg, Simon (2013), *Teacher Research in Language Teaching: A Critical Analysis*, Cambridge, Cambridge University Press.

Bowles, Melissa A. (2010), *The Think-Aloud Controversy in Language Acquisition Research*, New York, Routledge.

Chapelle, Carol A./Duff, Patricia. A. (2003), *Some Guidelines for Conducting Quantitative and Qualitative Research in TESOL*, TESOL Quarterly 37, 157–178.

Cohen, Jacob (^{2}1988), *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ, Erlbaum.

Davies, Mark (2002–), *Corpus del Español: 100 Million Words, 1200s–1900s*, http://www.corpusdelespanol.org. (10.10.2013)

Díaz-Campos, Manuel (2011), *Becoming a Member of the Speech Community: Learning Sociophonetic Variation in Child Language*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Malden, MA, Wiley-Blackwell, 263–282.

Dörnyei, Zoltán/Taguchi, Tatsuya (^{2}2009), *Questionnaires in Second Language Research: Construction, Administration and Processing*, New York, Routledge.

Ellis, Nick C./Sagarra, Nuria (2011), *Learned Attention in Adult Language Acquisition: A Replication and Generalization Study and Meta-analysis,* Studies in Second Language Acquisition 33, 589–624.

Ellis, Rod/Basturkmen, Helen/Loewen, Shawn (2001), *Preemptive Focus on Form in the ESL Classroom*, TESOL Quarterly 35/3, 407–432.

Erlam, Rosemary (2006), *Elicited Imitation as a Measure of L2 Implicit Knowledge: An Empirical Validation Study*, Applied Linguistics 27/3, 464–491.

Félix-Brasdefer, César (2008), *Perceptions of Refusals to Invitations: Exploring the Minds of Foreign Language Learners*, Language Awareness 17/3, 195–217.

Fernández, Claudia (2008), *Reexamining the Role of Explicit Information in Processing Instruction*, Studies in Second Language Acquisition 30, 277–305.

Foucault, Michel (1981), *The Order of Discourse*, in: Robert Young (ed.), *Untying the Text: A Poststructural Anthology*, Boston, Routledge, 48–78.

Friedman, Debra A. (2011), *How to Collect and Analyze Qualitative Research*, in: Alison Mackey/Susan M. Gass (edd.), *Research Methods in Second Language Acquisition: A Practical Guide*, Malden, MA, Wiley-Blackwell, 180–200.

Gass, Susan M./Mackey, Alison (2000), *Stimulated Recall Methodology in Second Language Research*, Mahwah, NJ, Erlbaum.

Glaser, Barney G./Strauss, Anselm L. (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Chicago, Aldine.

Golato, Andrea (2003), *Studying Complement Responses: A Comparison of DCTs and Recordings of Naturally Occurring Talk*, Applied Linguistics 24, 115–146.

Gurzynski-Weiss, Laura (2013), *Instructor Characteristics and Classroom-Based SLA of Spanish*, in: Kimberly L. Geeslin (ed.), *The Handbook of Spanish Second Language Acquisition*, Malden, MA, Wiley-Blackwell, 530–546.

Gurzynski-Weiss, Laura (in press), *Graduate Instructor In-class Cognition and Feedback Provision over Time*, in: Ryan T. Miller et al. (edd.), *Selected Proceedings of the 2012 Second Language Research Forum*, Somerville, MA, Cascadilla.

Harklau, Linda (2005), *Ethnography and Ethnographic Research on Second Language Teaching and Learning*, in: Eli Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, Mahwah, NJ, Erlbaum, 179–194.

Hashemi, Mohammad R./Babaii, Esmat (2013), *Mixed Methods Research: Toward New Research Designs in Applied Linguistics*, Modern Language Journal 97, 828–848.

Hernández, Todd (2006), *Integrative Motivation as a Predictor of Success in the Intermediate Foreign Language Classroom*, Foreign Language Annals 39/4, 605–617.

Jiang, Nan (2012), *Conducting Reaction Time Research in Second Language Studies*, New York, Routledge.

Labov, William (1972), *Sociolinguistic Patterns*, Philadelphia, University of Pennsylvania Press.

Larson-Hall, Jenifer/Herrington, Richard (2010), *Improving Data Analysis in Second Language Acquisition by Utilizing Modern Developments in Applied Statistics*, Applied Linguistics 31, 368–390.

Larson-Hall, Jennifer/Plonsky, Luke (forthcoming), *Reporting and Interpreting Quantitative Research Findings: What Gets Reported, How, and Why?*, in: John. M. Norris/Steven Ross/Rob Schoonen (edd.), *Improving and Extending Quantitative Reasoning in Second Language Research*, Malden, MA, Wiley.

Loewen, Shawn/Philp, Jenefer (2012), *Instructed Second Language Acquisition*, in: Alison Mackey/ Susan M. Gass (edd.), *Research Methods in Second Language Acquisition*, Malden, MA, Wiley-Blackwell, 53–73.

Loewen, Shawn, et al. (in press), *A Discipline Formed?: An Update on Applied Linguists' Statistical Literacy*, TESOL Quarterly.

Mackey, Alison/Gass, Susan M. (2005), *Second Language Research: Methodology and Design*, New York, Routledge.

Mackey, Alison/Gass, Susan M. (edd.) (2012), *Research Methods in Second Language Acquisition: A Practical Guide*, Malden, MA, Wiley-Blackwell.

Mackey, Alison, et al. (2010), *Exploring the Relationship Between Modified Output and Working Memory Capacity*, Language Learning 60/3, 501–533.

Norris, John M./Ortega, Lourdes (2006), *The Value and Practice of Research Synthesis for Language Learning and Teaching*, in: John M. Norris/Lourdes Ortega (edd.), *Synthesizing Research on Language Learning and Teaching*, Philadelphia, Benjamins, 3–50.

Ortega, Lourdes (1999), *Planning and Focus on Form in L2 Oral Performance*, Studies in Second Language Acquisition 21, 109–148.

Pavaresh, Vahid/Tavakoli, Mansoor (2009), *Discourse Completion Tasks as Elicitation Tools: How Convergent Are They?*, The Social Sciences 4/4, 366–373.

Plonsky, Luke (2011), *The Effectiveness of Second Language Strategy Instruction: A Meta-analysis*, Language Learning 61, 993–1038.

Plonsky, Luke (2012), *Effect size*, in: Peter Robinson (ed.), *The Routledge Encyclopedia of Second Language Acquisition*, New York, Routledge, 200–202.

Plonsky, Luke (2013), *Study Quality in SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research*, Studies in Second Language Acquisition 35, 655–687.

Plonsky, Luke (2014), *Study Quality in Quantitative L2 Research (1990–2010): A Methodological Synthesis and Call for Reform*, Modern Language Journal 98, 450–470.

Plonsky, Luke/Gass, Susan (2011), *Quantitative Research Methods, Study Quality, and Outcomes: The Case of Interaction Research*, Language Learning 61, 325–366.

Plonsky, Luke/Mills, Susana. V. (2006), *An Exploratory Study of Differing Perceptions of Error Correction Between a Teacher and Students: Bridging the Gap*, Applied Language Learning 16, 55–74.

Plonsky, Luke/Oswald, Frederick L. (forthcoming), *How Big is "Big"? Interpreting Effect Sizes in L2 Research*, Language Learning.

Poplack, Shana (2001), *Sometimes I'll Start a Sentence in Spanish y Termino en Español: Toward a Typology of Code-Switching*, in: Wei Li (ed.), *The Bilingualism Reader,* London, Routledge, 581–618.

Richards, Keith/Ross, Steven J./Seedhouse, Paul (2012), *Research Methods for Applied Language Studies: An Advanced Resource Book for Students*, New York, Routledge.

Schmidt, Richard/Frota, Sylvia Nagem (1986), *Developing Basic Conversational Ability in a Second Language: A Case Study of an Adult Learner of Portuguese*, in: Richard Day (ed.), *Talking to Learn: Conversation in Second Language Acquisition*, Rowley, MA, Newbury, 237–326.

Schmidt, Richard (1990), *The Role of Consciousness in Second Language Learning*, Applied Linguistics 11/2, 129–158.

Spada, Nina/Fröhlich, Maria (1995), *The Communicative Orientation of Language Teaching Observation Scheme (COLT)*, Sydney, MacMillan.

Tracy-Ventura, Nicole, et al. (2014), *"Repeat as Much as you Can": Elicited Imitation as a Measure of Oral Proficiency in L2 French*, in: Pascale Leclercq/Amanda Emonds/Heather Hilton (edd.), *Measuring L2 Proficiency: Perspectives from SLA*, Bristol, Multilingual Matters, 167–190.

Vinther, Thora (2002), *Elicited Imitation: A Brief Overview*, International Journal of Applied Linguistics 12/1, 54–73.