How to design and analyze language acquisition studies

Hugo Quené

Utrecht inst of Linguistics OTS, Utrecht University

h.quene@uu.nl

revision R5, 2010.04.01

1 Introduction

One of the key questions in linguistics is how language is acquired, both by children and by adults. Language acquisition is often investigated by means of behavioral research methods. The aim of the present chapter is to provide an overview of the most important methodological issues involved in designing empirical linguistic studies, and in analyzing data from such studies.

Solid research methods are not only required for good, and publishable scientific studies, but also for good ethics. This is because the effort for the human participants in terms of time, inconvenience, or loss of privacy should be outweighted by the expected scientific results (Rosnow & Rosenthal, 2001, Ch.3). Hence, if a study is unlikely to allow valid conclusions, then it would in general be ethically wrong to waste subjects' time, energy, and privacy,

and to expose them to unforeseen risks for this non-purpose.

More practically, researchers in many countries need to comply with legal regulations when human subjects are involved. In the U.S.A. this is enforced by a local Internal Review Board or Human Subjects Committee. The U.S. regulations simply state that the proposed study must "[use] procedures which are consistent with *sound research design*..." (Office of Human Research Protections, 2005, §46.111; emphasis added). The Dutch code of conduct for researchers requires them to exercise scrupulousness ("zorgvuldigheid") and requires that the benefits of the research should justify the risks for human (and animal) subjects (VSNU, 2004). Similar guidelines apply in other countries, and to all research funded through the E.U. (CORDIS, 2008). In general, such justification is possible only if the research is methodologically solid and adequate.

Methodological considerations are even more important if the subjects are children, in particular non-typically-developing children (see Chapters 11 and 12). Relative to adult participants in language research, children are more difficult to recruit, they are more vulnerable (hence the parents' informed consent is always required), they can perform a smaller range of tasks, and their attention and memory spans are shorter. Hence, special care is needed in designing a study, in recruiting and testing and protecting participants (see Editors' Introduction), and in analyzing their behavioral data. Consequently, it should go without saying that researchers should work out the study's design and data analysis in detail, before recruiting participants and collecting data.

2 Testing hypotheses

In empirical research, insights are primarily based on verifiable and objective observations, combined with logic, and not based on authority, common sense, or introspection (cf. Maxwell & Delaney, 2004; Rosnow & Rosenthal, 2001). Observations should also be consistent and reproducible, in order to obtain general insights from the limited sample of observations.

But how can such objective and reproducible observations lead to scientific insights? Let's consider the claim that all flames are hot. Does this claim gain empirical strength by finding positive evidence, i.e. by observing flames that are indeed hot? In fact, it does *not*, because of the so-called induction problem, already described by Hume (1739-1740, I.III.VI). Briefly, this problem entails that it is not logically safe to generalize from the observed cases to a general statement. Such generalizations always require a leap of faith from the observed instances to the general cause or principle. However, it is logically safe to refute the above claim by observing just one flame that is not hot, as was argued by Popper (1959/2002). Thus, falsification of the original claim has led to the insight that that claim was not correct, and we have gained a logically solid insight (that it is not true that all flames are hot).

Following this logic of falsification, a researcher typically studies two hypotheses. One hypothesis is related to the original research idea, e.g. the idea that the receptive vocabularies of 5-year-old children are larger than those of 3-year-old children. This is called the alternative hypothesis, or H_1. Positive evidence in favor of this H_1 is however not convincing, because of the induction problem mentioned above. (The positive evidence could come from biased selection of participants, or biased measurements, etc.) It would be more convincing if the researcher were to attempt to prove the logical

opposite of H_1; this opposite is often called the *null hypothesis*, or H_0. Here H_0 would claim that the 5-year-olds' vocabularies are not larger than those of the 3-year-old children. If the observations are very unlikely given the null hypothesis, then the researcher may reject the null hypothesis, and we may logically conclude that the alternative hypothesis or research idea is indeed correct.

Notice that if the observations are indeed likely, i.e. *not* unlikely, given the null hypothesis, then the researcher will not reject the null hypothesis. This does not imply that the null hypothesis is true, because "absence of evidence is not evidence of absence" (Sagan, 1996, p.221). A null result only implies that H_0 cannot be rejected convincingly. Hence there is an asymmetry: the null hypothesis is accepted by default, and only rejected by strong empirical evidence against it. In the present example, the older children's vocabularies may indeed be larger in reality than those of the younger children, but we may fail to observe this true difference in our sample, for a multitude of reasons. The null hypothesis would then not be rejected, even though it is in fact false (this is called a Type II error; see below for further discussion).

3 Types of studies

Empirical research attempts to find relations between variables. In the above L1 example, the main hypothesis claims that the variables children's age and vocabulary size are related. In second language acquisition, a researcher might hypothesize that learner proficiency is related to level of motivation. In a true *experiment*, the independent variable (also called explanatory variable) is manipulated by the experimenter, yielding different experimental conditions. Participants are randomly selected from the population(s) of

interest, and randomly assigned to these experimental conditions, in which their response values of the dependent variable are then observed. Hence, the observations of the dependent variable depend on the experimental conditions defined by the independent variable.

Many explanatory variables, however, cannot be manipulated at will by the investigator, because they constitute inherent properties of the individual participants. Examples are the participant's age, native language, clinical status, etc. For this reason, most language acquisition researchers use a quasi-experimental design, in which participants are *not* randomly assigned to experimental conditions. Such a *quasi-experiment* can succesfully establish relations between such independent variables and the dependent variable, although it may not be clear what is the cause and what the effect. For example, let us consider a fictitious observational study on the acquisition of a second language (henceforth L2), which shows that acquirers who have a higher motivation to learn the L2 also produce fewer errors in the L2. Does higher motivation cause higher proficiency, or is it the other way round? Or are both motivation and proficiency related to an unknown third variable, e.g. the amount of use of the L2?

In the latter example, the amount of use of the L2 may have been a confounding variable: a variable that is extraneous to the study, and not directly under investigation, but that is nevertheless related to both the independent and dependent variables. For example, amount of use may be related to proficiency because the more a learner uses a language, the more likely he is to become more proficient; amount of use could also be related to motivation because a more highly motivated learner may seek out more opportunities to use L2. More generally, a quasi-experimental study cannot entirely prove that the independent variable *causes* the observed effect in the dependent

variable. The direction of causality may be reversed, or as noted above the observed effect may be caused by other, confounding variables, which are not properly controlled because participants are not randomly assigned to conditions. Any conclusions about the causality of the observed relationship should therefore be drawn with caution, and only after considering possible confounding variables. If we need to be absolutely certain in identifying cause and effect, then a true experiment is required.

In the most basic experimental design, there is only one independent variable (usually categorical), and one dependent variable. The researcher samples groups of participants out of the population (e.g. two groups of children, of ages 3 and 5). The factor of interest, e.g. age, then varies "between subjects" or between groups of subjects. Such a cross-sectional design with different age groups may be used, for example, to assess whether 5-year-old and 7-year-old children process language differently (see Chapter 7). This design has the advantage that there is no transfer (e.g. learning) among conditions. The disadvantage however is that any accidental differences among the groups may be confounded with the main factor. Some confounding variables may be minimized by increasing the number of participants, with random sampling of participants from the population of interest. But other contextual confounds may be difficult to neutralize. In the above example, the 7-year-olds have lived through a longer and different history than the 5-year-olds, the older children are more developed, and they have a larger working memory and larger vocabulary than the younger children. All these differences may affect their performance. Similarly, when it comes to L2 acquisition, two groups differing in their native languages probably also differ in other relevant properties. Chapters 10 and 11 discuss how to minimize possible confounds in between-subject comparisons.

An alternative design is to vary a factor "within subjects", observing the same participant under multiple conditions, yielding "repeated measures" for each participant. This allows the researcher to disentangle the variation among participants from the effects of the main factor, yielding higher statistical power (see below). Consequently, fewer participants are needed in a within-subject study than in a between-subject study with equal power (Maxwell & Delaney, 2004, p.562). As a relevant example involving age, a linguistic researcher could draw a single sample of 5-year-olds, observe the participants' behavior, then wait 2 years, and observe the same participants again at age 7. In this so-called "longitudinal" design, participants may have transferred experiences from previous to subsequent observations. For example, they may have learned how to perform in language tests. They may also dropped out of the study in a non-random fashion (e.g. because of fast or slow rate of development). Hence the main factor may be confounded with other developmental and external variables. Longitudinal designs are used to investigate language acquisition, e.g. in diary studies, but they require considerable time and effort from the researcher and the participants (see Chapter 1).

4 Validity

Any scientific study aims to obtain valid insights from empirical observations. Valid conclusions are only justified, however, if the study was properly designed, conducted and analyzed. The term *validity* refers to how correct or accurate the conclusions of a scientific study are (Maxwell & Delaney, 2004).

As we saw above, this validity is threatened by nuisance variables and confounds, which may provide plausible alternative explanations for an ob-

served effect. Experimental designs may be ranked by their susceptibility to such threats of validity. At one end we find a particularly strong design (randomized true experiments, typically used in medical and pharmaceutical research), intermediate positions are taken by other designs (e.g. quasi-experiments), and at the other end we find designs which are very weak when it comes to validity (e.g. uncontrolled case studies, which may nevertheless provide useful insights).

Validity may be threatened by contextual factors already mentioned, including maturation of participants, and artefacts introduced by our measuring instruments such as verbal or cognitive tests (e.g. Rosenthal & Rosnow, 2008, p.211). These possible confounds may be controlled in longitudinal research by including a control group for comparison. Methods to control confounding variables are discussed in Chapters 2 (for adult participants), 10 (for children and adults) and 11 (for language impaired children).

In addition, validity in acquisition research may be threatened by selection bias, in particular volunteer bias (Rosenthal & Rosnow, 1969). It is quite plausible that people who are relatively more likely to volunteer themselves (or their children) for language research, also have better than average linguistic and verbal skills. This is obvious in many diary studies, and in the Childes database (MacWhinney, 2000), where the proportion of children with highly educated parents is far larger than in the general population. The children of highly educated parents may well have had more intense and more focused language exposure (perhaps combined with an inherited verbal giftedness) than other children. As a result, the diaries and databases are not necessarily representative of the language populations they attempt to represent. Hence caution is required in generalizing findings from such biased samples to a wider population of language acquirers.

5 Significance, power and effect size

Let us return now to the logic of testing hypotheses by means of empirical data, temporarily ignoring the experimental design, and focusing on the data analysis. As explained above, the null hypothesis to be tested often states that the true effect (or difference) in the population is absent. If this H_0 is true, then the observed effect in the sample is likely to be very small as well — but due to sampling variation, a larger effect may be observed occasionally. Statistical analysis tells us how likely it is to find the observed effect, or a more extreme one, if H_0 is true.

If the probability (abbreviated as p) of the observed effect is very low given H_0, then this may be regarded as convincing or "significant" evidence against that H_0. The basic argument is as follows: an effect has been observed; if H_0 is true then this effect is very unlikely; therefore H_0 is rejected and the alternative H_1 is accepted. The level of *significance*, or probability p under H_0, is also the risk of rejecting H_0 incorrectly, i.e., of finding an effect even if H_0 is true (a false positive). This incorrect rejection of H_0 is called a Type I error. The cut-off value for p, i.e. the highest acceptable risk of incorrectly rejecting H_0, is called α; an often-used cut-off value is $\alpha = .05$. In sum, significance refers to the probability of observing this effect (or a larger effect) given H_0, and *not* to the probability of H_0 given the observed effect (Cohen, 1990, 1997).

The risk of committing a Type I error (of regarding a null effect as significant) should be balanced against the opposite error of Type II, of failure to regard a non-null effect as significant (a miss, or false negative). This error occurs if we fail to reject H_0 even though H_0 is in fact false. The risk of this Type II error is indicated by β. If H_0 is indeed false, then the complement of this risk, or $1-\beta$, constitutes what is called the statistical *power* of the study.

The power is the probability of rejecting H_0 if H_0 is indeed false. Informally speaking, this is the chance of corroborating your H_1 if H_1 is indeed true, or the chance of you getting it right if you are right (a hit, or true positive). Hence the risk of a Type I error (α) should be balanced against the risk of a Type II error (β). Many studies are conducted with maximum error probabilities of $\alpha = .05$ and $\beta = .20$ (power .80). Hence in these studies a Type I error is — admittedly somewhat arbitrarily — regarded as four times as costly than a Type II error. If we regard both errors as more equally serious, however, then we might better use a higher α, and/or lower β, or both (Rosenthal & Rosnow, 2008, Ch.12).

More important than the binary decision regarding H_0 is the size of the hypothesized effect. Even a very small effect may be statistically significant if the size of study or number of observations is very large; this is succinctly summarized as "significance test = size of effect $\times$ size of study" (Rosenthal & Rosnow, 2008). For example, the small difference in vocabulary size between children aged 5;0 and 5;1 will be statistically significant if we include hundreds of participants in each age group. But we do not want to spend large research funds, and waste many participants' time and effort, only to report vanishingly small and irrelevant effects as significant. This means that we should think about the smallest effect that we consider relevant and that we wish to detect in our study (see below for further discussion). Moreover, we should habitually report not only the significance level, but also the size of the observed effect. This is part of the research guidelines implemented by some scientific journals, e.g. *Language Learning* and *TESOL Quarterly*. In our example study on vocabulary size by children of ages 3 and 5 (§2), differences smaller than 2 scoring units might be regarded as irrelevant; the smallest relevant difference is 2 units. The study should be designed such

that the statistical power is at least .8 (or $\beta \leq .2$) for detecting a vocabulary size difference of 2 units or larger.

Multiple observations yield different outcomes (otherwise research would be quite boring). The amount of dispersion among observations is called the *standard deviation* (symbol s). This dispersion may be due in part to random fluctuations, to irrelevant individual differences among participants, and to measurement errors. The effect we are investigating is hidden, as it were, in this random dispersion among observations. As you can imagine, detecting a small effect in a set of observations is easier if the random dispersion among observations is relatively low. Hence the effect under investigation is expressed relative to the standard deviation. The resulting relative effect size (symbol d, effect divided by dispersion) thus indicates the contrast or relative conspicuousness of the hypothesized effect against the random variability among observations (Cohen, 1988). If the hypothesized effect yields a difference of 2 scoring units, and the dispersion s among observations is 4 scoring units, then the relative effect size d is 2/4 or 0.5. If the hypothesized difference is only 1 scoring unit, and the dispersion is 2 scoring units, then the relative effect size is 1/2 or 0.5 also. By contrast, if the hypothesized difference is 4 scoring units, but the dispersion is as large as 20 scoring units, then the relative effect size is only 4/20 or 0.2. If the random dispersion or variability among observations is smaller, then an effect is more likely to be detected; in other words, statistical power increases. Hence, it is worthwile for researchers to think about methods to reduce random variability. The relations between significance, power, sample size, standard deviation, and effect size (Cohen, 1988; Lipsey, 1990; Rosenthal, Rosnow, & Rubin, 2000; Rosenthal & Rosnow, 2008) will be further illustrated below.

Why should researchers worry about power? The first reason is of course

Type II error itself, which may have immediate and possibly serious consequences. But there are methodological and practical considerations, too. Let us consider the example of a study in which one group of participants (e.g. bilinguals) is hypothesized to perform better than the other group (e.g. monolinguals), on some dependent variable reflecting linguistic performance (see Bialystok, 2001). Let us also assume that H_0 is indeed false, so the two types of speakers indeed perform differently; this should yield a significant group effect on linguistic performance. If several replicated low-power studies are taken together, then a significant effect may be found in some studies, but not in others, due to the low power in each study. Subsequently, researchers typically attempt to explain these different outcomes in a series of follow-up studies.

Researchers should realize, however, that a mix of significant and non-significant findings may well be due to the low power in each study, and *not* necessarily to other differences in the studies, e.g. differences in stimulus materials, testing procedure, properties of the participants, etc. Focusing on these differences between low-power studies may easily lead "to wasted research efforts to identify nonexistent moderator variables" (Schmidt, 1996, p.118).

Moreover, the scientific record of published studies will contain a confusing mix of significant and non-significant findings. Many professionals rely on this scientific record for their work, e.g. for developing diagnostic tools and evidence-based treatment programs in education and in health care, or for further scientific research. The mixed outcomes of the multiple studies will prevent these professional consumers from concluding that the two groups of participants (in the current example) do indeed perform differently (Van Kolfschooten, 1993, p.92). This may seriously hamper progress

in diagnosis, treatment, and research.

6 Frequently asked questions

Following the basics of experimental design and hypothesis testing, this section addresses some frequently asked questions about these topics when it comes to conducting research into language acquisition. It attempts to explain why various properties of your study are important in the answers to these questions. We start with the most frequently asked question.

6.1 How many participants and items are required?

In order to answer this question, we need to take other properties of the study into account. The minimum number of participants (and items) depends on the chosen level of significance (α), on the desired power ($1 - \beta$), and on the expected relative effect size (d, expected difference divided by expected standard deviation). These concepts were introduced in §5 above.

In order to illustrate the complex relations among these key properties of a study, let us regard a fictitious head-turn preference study. A sample of infants is compared on their listening time (in seconds), under two conditions using a within-subject design (after the Modified Head-Turn Preference procedure as used by Jusczyk & Aslin, 1995; see also Chapter 4, section 3, of this volume). In one condition the participants listened to target words after a preceding period of familiarization, and in the other condition they listened to other targets without such a period of familiarization. The research hypothesis H_1 states that there is an effect of familiarization, i.e., a difference among the two conditions. Because shorter listening times are expected after familiarization, the expected effect is negative, i.e., a decrease in listening

time due to familiarization. The corresponding null hypothesis states that this effect of familiarization is absent, or nil. We assume conventional criteria of $\alpha = .05$ for a Type I error, and $\beta = .20$ for a Type II error (power .80). We also assume that the smallest difference of interest is 1 second (this is the difference in listening time between experimental conditions). With dispersion assumed to be $s = 2$ seconds, this amounts to an expected relative effect size of $d = 1/2$. Stated differently, we want to have at least 80% chance (power .80) of detecting an effect of size 1/2 (difference of 1 divided by variability of 2 seconds), and we also want to have at most 5% chance of erroneously reporting an effect that is in fact nil. As you might imagine, the researcher's task of discriminating a relevant effect from irrelevant variability becomes easier as more participants (and items) are included in the study. But how many are sufficient?

In this example, if we still assume a within-subject design, then a conservative estimate for the minimum number of participants is $n = (2.8/d)^2$, rounded up to 32 participants (Winer, 1971; Cohen, 1988; Lenth, 2006; Gelman & Hill, 2007). If we would assume a between-subject design, and all other properties unchanged, then at least $n = (5.6/d)^2$, rounded up to 126 participants would be required (Gelman & Hill, 2007, Ch.20), because between-subjects designs are less efficient and require more participants to obtain the same power. In either design, the number of required participants is larger as the relative effect size is smaller. If the expected relative effect size is halved to $d = 1/4$, then $n = 126$ participants are required in a within-subject design, and $n = 502$ participants are required in a between-subject design. Obviously, detecting a relatively small effect requires relatively many participants and items, and vice versa. (The fixed values of 2.8 in the first and 5.6 in the second formula above capture the combined values of α and

β, taking design properties into account, Gelman & Hill, 2007).

Thus, in order to determine the number of participants (and items), the researcher needs to know the essential statistical properties of the study. First, a rough estimate is required of the expected difference due to experimental groups or conditions. Second, an estimate of the standard deviation is required; this is often derived from previous studies. If no previous studies are available, the standard deviation may be estimated from a dozen or so pilot observations: take the highest and lowest observation, compute their difference (called the *range*), and divide the range by 4, yielding a rough estimate of the standard deviation (Peck & Devore, 2008, p.399). Third and fourth, appropriate risk levels for Type I and Type II errors must also be determined, as explained above. Finally, the experimental design needs to be chosen. After all these essential properties of the study are determined, the researcher may use formulae (e.g. Gelman & Hill, 2007) or dedicated software (e.g. Lenth, 2006) to compute the required minimum number of observations. It is better to include a few more participants than this required minimum.

6.2 What if only a small number of participants are available?

In many situations, recruiting many participants (or constructing many test items) is not possible, and researchers will have to compromise in sample sizes. As explained above, this reduces the power or sensitivity of the study. A non-significant outcome, hence failure to reject H_0, could either be due to H_0 being true, or it could be due to a Type II error. This implies that no conclusions should be based on a null result, if the observed power in detecting a relevant effect was low, as discussed in section 2 above. Caution

is required even if H_0 was rejected (i.e., a significant outcome), because results from a small sample may not generalize to the population. A smaller sample also has a smaller chance of being representative of the population from which it was drawn.

We saw in the preceding section that the required number of participants depends in part on the relative effect size d (relevant difference divided by irrelevant variability). So, one could compensate for the smaller sample size by sacrificing sensitivity, i.e., by sacrificing the power to detect small effects. A larger relevant difference, and unchanged variability, will yield a larger relative effect size d. One could also attempt to reduce the random variability, as will be discussed in the following section. In the fictitious head-turn preference study discussed above, for example, power could be maintained at about .75 with only $n = 9$ participants, *if* the smallest detectable difference in listening time is increased from 1 to 2 seconds, meaning that the smallest detectable relative effect size d is increased from $1/2$ to $2/2 = 1$.

With this few participants, however, only large effects can be detected reliably, and medium-sized effects are most likely to go undetected, even though it might be such a medium-sized effect in which the researcher is interested. If the effect is medium-sized (e.g. $d = 1/2$), and if only 9 participants are available, then the observed power would be as low as .26, far below the desired level of .80.

Thus, if the required number of observations (participants and/or test items) cannot be obtained in a study, then a researcher should accept that only large effects can be assessed, and that small effects cannot be assessed. In addition, failure to find a significant effect (i.e., failure to reject H_0) does not imply that the hypothesized effect is nil (see section 2). It is therefore informative to report and discuss the relative effect size, as discussed above.

6.3 How can I increase the sensitivity of my study?

As explained above, the sensitivity of testing hypotheses can be maintained to some extent by increasing the relative effect size, if the number of participants and/or items is low. In addition to raising the bar for a relevant difference, a researcher might attempt to reduce the random variability among the observations. An unchanged relevant difference, and lower variability, will yield a larger relative effect size d. In the head-turn preference example discussed above, power could be maintained at about .75 with only $n = 9$ participants, and with a minimum effect of 1 second difference in listening time, *if* the random variability or standard deviation is decreased to $s = 1$ second. This increases the smallest detectable relative effect size d from $1/2$ to $1/1 = 1$.

So how can researchers reduce the standard deviation in their observations? While designing, conducting and analyzing a study, all possible sources of variation, except for the variance due to the construct under investigation, must be eliminated as much as possible.

One important type of variation is that between individual participants, as extensively discussed in the three chapters of this volume that deal with comparisons across groups (Chapters 10, 11, 12). Variation between individual participants can be addressed during recruitment (sampling of participants), by including participants that form as homogeneous a group as possible while still representative of (and randomly selected from) some reference population (Lipsey, 1990). If we study language acquisition in bilingual children, for example, then participants may be difficult to recruit. One would be tempted to include all available children, irrespective of their parents' native languages. In fact, it might be wiser to select children by keeping the parents' languages fixed throughout the sample (e.g. all selected children have a father with native language X and mother with native language Y).

This may yield fewer children in the study, and it may reduce the generalizability of its results to a smaller population of interest, but the smaller and more homogeneous sample is also likely to yield smaller variation between observations, and hence a larger probability of a positive outcome.

In designing a between-subjects study, especially with few participants, it may also be worthwile to *match* participants from the various groups on relevant confounding variables (such as gender, socio-economic status, age), rather than to rely on random selection to cancel out these confounds (Moore, McCabe, & Craig, 2009, Ch.3). Specific suggestions on how to match and to reduce variation between individual participants can be found in the aforementioned chapters on comparing groups.

During the test itself, random variability in the measurements can be reduced by using a protocol for testing and quantifying observations (Lipsey, 1990), spelling out the procedures for the experimenters on how to instruct participants, take the tests, make and record observations, transcribe and quantify responses, etc. Such instructions will ultimately reduce standard deviations, e.g., because all transcribers follow the same instructions in similar cases. See the Editors' Introduction (Chapter 1) for further discussion of test protocols and of methods to reduce variability.

The purpose of these different ways to reduce variability is to reduce the random variability among observations. This increases the relative effect size d, and this in turn provides some compensation against low numbers of participants and/or items. If our alternate research idea is indeed true, the ultimate aim of these efforts is to protect the power or sensitivity of our study as much as possible. High power is desirable, after all, because we would like to have a high chance of rejecting a false H_0.

6.4 How can I prove that there is no difference?

The logic of falsificationism, together with the asymmetry among H_1 and H_0, brings problems if one attempts to *verify* a H_0 which states that two constructs are identical (that there is no effect). The opposite hypothesis claims that responses in the two conditions are different, and predicts that there will be some difference between conditions. And indeed, such differences will *always* be observed in real life, if only because of sampling variations, so that a no-effect hypothesis can never be verified (it can only be falsified). Unfortunately, many linguistic studies run into this problem of "proving the null hypothesis", because the aim of these studies is to investigate the similitudes among languages and in language behavior under different circumstances or by different groups of participants. Many of the studies discussed in Chapter 10, for example, aim to show that there is no fundamental difference between the ways in which children and adults acquire grammar.

One sensible solution for this problem is to acknowledge that H_1 and H_0 are effectively reversed here, and so the conventional risks of Type I and Type II errors should also be reversed. The H_0 which we attempt to verify should be rejected relatively easily, say $\alpha = .20$, and the power in detecting a small-size effect should be high, say $\beta \leq .05$, to fend off the critique that we have not attempted strong enough to reject H_0.

In one of the fictitious example studies mentioned above (§2), vocabulary sizes of 3-year-old and 5-year-old childen were compared, with H_0 stating that there is no difference between these two groups in vocabulary size. *If* the statistical power were really high, say above 95% for differences of 2 scoring units or larger, and if H_0 can nevertheless not be rejected, then this might be interpreted as evidence that H_0 is probably true. The chance of this conclusion being incorrect (a Type II error) is then below 5%.

The margin of error of this decision is equivalent to the effect size discussed in §5. In conventional null-hypothesis significance testing, the effect size is like the weakest relevant effect (the weakest signal) that we want to detect in the irrelevant background variability (the noise), in order to reject H_0. In the reverse procedure, the margin of error is like the highest noise level that we want to allow in finding silence, in order to accept H_0.

The reverse procedure sketched above may allow us to verify the null hypothesis, but unfortunately it requires many participants. In this example study, having a between-subjects design, the required number of participants is at least $n = 138$ children (see §6.1; Gelman & Hill, 2007, Ch.20), or 69 children in each group. Thus we can only verify a null-effect hypothesis with some confidence if we use large samples, or large margins of error, or both (Cohen, 1988).

6.5 What to do about missing data?

Missing data are a cause for concern, for two reasons. First, the lower number of observations reduces statistical power. Secondly, observations are often not missing at random but according to some pattern, which introduces bias in the remaining observations, and hence threatens validity. The longitudinal example study comparing performance at ages 3 and 5 (§3) may be biased, if the drop-out pattern is somehow related to the participants' linguistic performance. The pattern of missingness should be inspected as a regular part of the data analysis, because it can reveal interesting properties of the participants' behavior as well as identify possible biases in the remaining data.

Statistical analyses that are based on comparisons (e.g. t test, analysis of variance) typically require complete cases, so the number of remaining com-

plete cases (participants) for analysis may be quite low. (Cases are deleted "listwise", i.e. if just a single observation is missing.) In general, statistical analyses that are based on regression (e.g. linear or logistic regression, multiple regression, mixed effects modeling, or factor analysis) are far more robust against missing data. It may be worthwile therefore to analyze the data by means of a regression-based technique. This can be achieved by coding the independent variables as dummy predictors, and including these in a multiple regression model (for analysis procedures, see e.g. Field, 2009, Ch.7). Random factor(s) of participants (and items) may also be included in such models. This yields a so-called *mixed effects* or *multi-level* model, which is often more adequate than a conventional analysis of variance or t test (for recent overviews, see e.g. Gelman & Hill, 2007; Quené & Van den Bergh, 2008; Baayen, 2008; Baayen, Davidson, & Bates, 2008; Field, 2009, Ch.19).

7 Dos and don'ts

- *Do* consider the design and analysis of a study together (see e.g. Levin, 1999; Kirk, 1995). You should really think about how the data will be analyzed *before* the data are collected, i.e., while designing your study.
- *Do* consider the best practices and recommendations given by experienced researchers (e.g. Cohen, 1990; Schmidt, 1996; Wilkinson & Task Force on Statistical Inference (APA Board of Scientific Affairs), 1999; Maxwell & Delaney, 2004; Rosenthal & Rosnow, 2008).
- *Do* reflect on the various possible threats to validity, and on how these may be neutralized effectively.

- *Do* learn about statistical analyses (and software) from many excellent statistical textbooks, e.g. StatSoft, Inc. (2007); Moore et al. (2009), Field (2009, for SPSS), Baayen (2008, for R) and Johnson (2008, for R).

- *Do* conduct pilot work before conducting your main experiment, in order to smooth out infelicitous choices in stimulus materials and test procedures, to estimate standard deviations (needed for power analysis), and to dry-run your data analysis.

- *Do not* include too many variables in your study, and *do* include more participants (and/or items) (Cohen, 1990). This is a variant of the well-known advice for travelers to pack half the clothes, and double the money. Research progresses best in small steps, so that fewer things can go wrong, and fewer or smaller confounds may be distorting the hypothesized effect of interest.

- *Do* inspect whether there is some pattern in the missingness of observations, and whether you can (partly) account for this pattern. Learn more about methods to adjust your analyses, or to impute the missing data (e.g. Little & Rubin, 1987).

- *Do not* forget the hypotheses and ultimate objectives of your study while analyzing your data. Often the binary decision whether or not to reject the null hypothesis, at an arbitrary level of significance, does insufficient justice to your findings (Schmidt, 1996).

- *Do* make it a habit to report effect sizes and/or confidence intervals of your findings (following e.g. *Publication Manual of the American Psychological Association*, 2001, p.26). These indicators convey the *degree*

to which the null hypothesis is or is not true, whereas the above binary decision does not. Fellow researchers may use your reported effect sizes, e.g. for properly designing their own studies, and for assessing the magnitude of your hypothesized effect.

- *Do* use your own critical "informed judgment" (Cohen, 1990) as a behavioral linguistic researcher, during all stages of designing, conducting and analyzing a study. Realize that the perfect study still has to be conducted, even after you have finished yours. Your linguistic and methodological expertise are essential to arrive at valid and interesting conclusions about language acquisition.

Acknowledgement

My sincere thanks are due to Huub van den Bergh, Esther Janse, Sieb Nooteboom, Karin Wagenaar, the editors and all reviewers, for their helpful suggestions and comments on previous versions of this chapter.

References

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R., Davidson, D., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition.* Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cohen, J. (1997). The earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance test?* (pp. 21–35). Mahwah, NJ: Lawrence Erlbaum.

CORDIS. (2008). *Getting through ethics review.* Available: `http://cordis.europa.eu/fp7/ethics_en.html`.

Field, A. (2009). *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)* (3rd ed.). Los Angeles: Sage.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Hume, D. (1739-1740). *A Treatise on Human Nature.* Available: `http://www.gutenberg.org/dirs/etext03/trthn10.txt`.

Johnson, K. (2008). *Quantitative methods in linguistics.* Malden, MA: Blackwell.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1-23.

Kirk, R. (1995). *Experimental Design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Lenth, R. (2006). *Java applets for power and sample size.* Available: `http://www.cs.uiowa.edu/~rlenth/Power`.

Levin, I. P. (1999). *Relating statistics and experimental design: An introduction.* Thousand Oaks, CA: Sage.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database). Mahwah, NJ: Lawrence Erlbaum.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics* (6th ed.). New York: Freeman.

Office of Human Research Protections. (2005). *Protection of Human Subjects.* (Code of Federal Regulations, Title 45 Part 46.)

Peck, R., & Devore, J. (2008). *Statistics: The exploration and analysis of data* (6th ed.). Belmont, CA: Thomsom/Cole.

Popper, K. (1959/2002). *The logic of scientific discovery.* London: Routledge.

Publication Manual of the American Psychological Association (5th ed.). (2001). Washington, DC: American Psychological Association.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425.

Rosenthal, R., & Rosnow, R. L. (1969). The volunteer subject. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 59–118). New York: Academic Press.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). Boston: McGraw Hill.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge:

Cambridge University Press.

Rosnow, R., & Rosenthal, R. (2001). *Beginning Behavioral Research: A conceptual primer* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Sagan, C. (1996). *The demon-haunted world: Science as a candle in the dark*. Random House.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129.

StatSoft, Inc. (2007). *Electronic statistics textbook.* Tulsa, OK: StatSoft, Inc. Available: `http://www.statsoft.com/textbook/stathome.html`.

Van Kolfschooten, F. (1993). *Valse vooruitgang: Bedrog in de Nederlandse wetenschap*. Amsterdam: L.J. Veen.

VSNU. (2004). *Nederlandse Gedragscode Wetenschapsbeoefening.*

Wilkinson, L., & Task Force on Statistical Inference (APA Board of Scientific Affairs). (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Winer, B. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.