Chapter 16

ANALYSIS, INTERPRETATION, AND VISUAL PRESENTATION OF EXPERIMENTAL DATA

Geoffrey R. Loftus, *University of Washington*

Following data collection from some experiment, there arise two goals which should guide subsequent data analysis and data presentation. The first goal is for the data collector him or herself to understand the data as thoroughly as possible in terms of (1) how they may bear on the specific question that the experiment was designed to address, (2) what if any surprises the data may have produced, and (3) what, if anything, such surprises may imply about the original questions, related questions, or anything else. The second goal is to determine how to present the data to the scientific community in a manner that is as clear, complete, and intuitively compelling as possible. This second goal is intimately entwined with the first: Whatever data analysis and presentation techniques best instill understanding in the investigator to begin with are generally also optimal for conveying the data's meaning to the data's eventual consumers.

So what *are* these data-analysis and data-presentation techniques? It is not possible in a single chapter or even in a very long book to describe them all, because there are an infinite number of them. Although most practicing scientists are equipped with a conceptual foundation with respect to the basic tools of data analysis and data presentation, such a foundation is far from sufficient: it is akin to an artist's foundation in the tools of color mixing, setting up an easel, understanding perspective, and the like. To build on this analogy, a scientist analyzing any given experiment is like an artist rendering a work of art: Ideally the tools comprising the practitioner's foundation should be used creatively rather than dogmatically to produce a final result that is beautiful, elegant, and interesting, instead of ugly, convoluted, and prosaic.

My goal in this chapter is to try to demonstrate how a number of data-analysis techniques may be

The writing of this chapter was supported by NIMH grant MH41637 to G. Loftus. I thank the late Merrill Carlsmith for introducing me to numerous of the techniques described in this chapter and David Krantz for a great deal of more recent conceptual enlightenment about some of the more subtle aspects of hypothesis-testing, confidence intervals and planned comparisons.

used creatively in an effort to understand and convey to others the meaning and relevance of a data set. It is not my intent to go over territory that is traditionally covered in statistics texts. Rather, I have chosen to focus on a limited, but powerful, arsenal of techniques and associated issues that are related to, but are not typically part of a standard statistics curriculum. I begin this chapter with an overview of data analysis as generically carried out in psychology, accompanied by a critique of some standard procedures and assumptions, with particular emphasis on a critique of null hypothesis significance testing (NHST). Next, I discuss a collection of topics that represent some supplements and/or alternatives to the kinds of standard analysis procedures about which I will have just complained. These discussion include (1) a description of various types of pictorial representations of data, (2) an overview of the use of confidence intervals which, I believe, constitutes an attractive alternative to NHST, (3) a review of the benefits of planned comparisons which entail an analysis of percent between-conditions variance accounted for, (4) a description of techniques involving percent *total* variance accounted for, (5) a brief set of suggestions about presentation of results based on mathematical models (meant to complement the material in the Myung & Pitt chapter) and finally, (6) a somewhat evangelical description of what I have termed *equivalence techniques*.

My main expositional strategy is to illustrate through example. In most instances, I have invented experiments and associated data to use in the examples. This strategy has the disadvantage that it is somewhat divorced from the real world of psychological data, but has the dominating advantage that the examples can be tailored specifically to the illustration of particular points.

The logic and mathematical analysis in this chapter is not meant to be formal or complete. For proofs of various mathematical assertions that I make, it is necessary to consult a mathematically oriented statistics text. There are a number of such texts; my personal favorite is Hays (1973), and where appropriate, I supply references to Hays along with specific page numbers.

My choice of material and the recommendations that I selected to include in this chapter have been strongly influenced by 35 years of experience in reviewing and journal editing. In the course of these endeavors I have noticed an enormous number of data-analysis and data-presentation techniques that have been sadly inimical to insight and clarity—and conversely, I have noticed enormous numbers of missed opportunities to analyze and present data in such a way that the relevance and importance of the findings are underscored and clearly conveyed to the intended recipients. Somewhere in this chapter is an answer to approximately 70% of these complaints. It is my hope that, among other things, this chapter will provide a reference to which I can guide authors whose future work passes across my desk—as an alternative, that is, to trying to solve what I believe to be the world's data analysis and presentation problems one manuscript at a time.

FOUNDATIONS: THE LINEAR MODEL AND NULL HYPOTHESIS SIGNIFICANCE TESTING

Suppose that a memory researcher were interested in how stimulus presentation time affects memory for a list of words as measured in a free-recall paradigm. In a hypothetical experiment to answer this question, the investigator might select J = 5 presentation times consisting of 0.5, 1.0, 2.0, 4.0, and 8.0 sec/word and carry out an experiment using a between-subjects design in which n = 20 subjects are assigned to each of the 5 word-duration conditions—hence, N = 100 subjects in all. Each subject sees 20 words, randomly selected from a very large pool of words. For each subject, the words are presented sequentially on a computer screen, each word presented for its appropriate duration. Immediately following presentation of the last word, the subject attempts to write down as many of the words as possible. The investigator then calculates the proportion correct number of words (out of the 20 possible) for each subject.

The results of this experiment therefore consist of 100 numbers: one for each of the 100 subjects. How are these 100 numbers to be treated in order to address the original question of how memory performance is affected by presentation time? There are two steps to this data-interpretation process. The first is the specification of a *mathematical* model[1], within the context of which each subject's experimentally observed number results from assumed events occurring within the subject. There are an infinite number of ways to formulate such a mathematical model. The most widely used formulation, on which I will focus on in this chapter, is referred to as the *linear model* or LM.

The second step in data interpretation is to carry out a process by which the mathematical model, once specified, is used to answer the ques-

[1] I have sometimes observed that the term "mathematical model" casts fear into the hearts of many researchers. However, if it is numbers from an experiment that are to be accounted for, then the necessity of some kind of mathematical model is logically inevitable.

Table 1. Types of Models. The response measure is R, and the values of independent variables are labeled X and Y. The model parameters are indicated by Greek letters, α, β, γ, and δ. All models listed are linear models except for the last which is not linear because it includes the product of three parameters, $\delta\beta_i\gamma_j$.

Model	Model Name
$R = \alpha + \beta X + \gamma Y$	Multiple Regression (Additive)
$R = \alpha + \beta X + \gamma Y + \delta XY$	Multiple Regression (Bilinear)
$R = \alpha + \beta X + \gamma Y + \delta Y^2$	Multiple Regression (Quadratic in Y)
$R = \alpha + \beta_i + \gamma_j$	Two-Way ANOVA (Additive)
$R = \alpha + \beta_i + \gamma_j + \delta_{ij}$	Two-Way ANOVA with Interaction
$R = \alpha + \beta X + \gamma_j$	One-Way ANACOVA (Additive)
$R = \alpha + \beta_i + \gamma_j + \delta\beta_i\gamma_j$	Tukey's one-degree-of-freedom interaction model

tion at hand. Note that there are numerous possibilities for how this can be done. The process that is the most widely used is that of null hypothesis significance testing, or NHST.

Most readers of this chapter are probably familiar with both the LM and the process of NHST. Nonetheless, to insure a common conceptual and notational foundation, I will describe both of them briefly in the next two sections.

The Linear Model

The LM, although central to most statistical analysis is described by surprisingly few introductory statistics books (Hays, 1973, my statistics reference of choice in this chapter is one of them). The LM includes a variety of assumptions, the exact configuration of which depends on the nature of the experimental design. At its most general level, within the context of the Linear Model, some response variable, R, is modeled as a linear function of various parameters, labeled α, β, γ, δ, and so on. Table 1 provides some examples of common linear models along with the names of these models. For comparison purposes, the last entry in Table 1 is an example of a nonlinear model wherein one term is a product of several of the parameters. It is noteworthy, incidentally that (unlike many social science statistics texts and statistics courses) the Linear Model does not make a sharp distinction between ANOVA and regression. Instead, both are simply viewed as instances of the same general model.

In the simple free-recall example described above, the Linear Model is formulated thusly.

1. The subjects in the experiment are assumed constitute a random sample from some population to which conclusions are to apply.
2. Similarly, he words provided to each subject are assumed to be a random sample drawn from a large population of words.
3. Across the subjects x words population there is a "grand mean," denoted μ, of the dependent variable measured in the experiment. The grand mean is a theoretical entity, but roughly it can be construed as the number that would result if all individuals in the target population were run in the experiment for an infinite number of times in all conditions, using in the course of this lengthy process, the entire population of words, and the mean of all the resulting scores were computed.
4. Each condition, j, in the experiment has associated with it an "effect" which is referred to as α_j. Any score obtained by a subject in condition j is increased by α_j compared to the grand mean, μ. Over the population, the mean score for condition j, which is referred to as μ_j, is $\mu_j = \mu + \alpha_j$. The model defines these effects such that

$$\sum_{j=1}^{J} \alpha_j = 0$$

   which means, of course, that either all the α_j's are zero, or that some are positive while others are negative.
5. Associated with each subject participating in the experiment is an "error term" that is specific to that subject. This error term is independent of condition, and the error term for Subject i in Condition j is labeled e_{ij}. It is assumed that the e_{ij}'s are randomly drawn from a normal distribution whose mean is zero and whose variance is σ^2, a value that is constant over conditions[2].

[2] A technical point is in order here. The error term for this experiment has two components. The first is a subject component reflecting the fact that proportion correct varies among subjects, The second is a binomial component reflecting variation over the 20 words. Because the binomial variance component changes

These assumptions imply that the X_{ij}, the score of Subject i in Condition j is equal to,

$$X_{ij} = \mu + \alpha_j + e_{ij}$$

which in turn implies that X_{ij}'s within each condition j are distributed with a variance of σ^2.

Null Hypothesis Significance Testing

Equipped with a mathematical model, the investigator's next step in the data-analysis process is to use the model to arrive at answers to the question at hand. As noted, the most pervasive means by which this is done is via NHST, which works as follows.

1. A "null hypothesis" (H_0) is established. Technically, a null hypothesis is any hypothesis that specifies quantitative values for all the α_j's. In practice however a null hypothesis almost always specifies that "the independent variable has no effect on the dependent variable," which means that,

   $$H_0: \alpha_1 = \alpha_2 = ... = \alpha_J = 0$$

   or equivalently, that,

   $$H_0: \mu_1 = \mu_2 = ... = \mu_J$$

   Mathematically the null hypothesis may be viewed a single-dimensional hypothesis: the only variation permissible is the single value of the J population means.
2. An "alternative hypothesis" (H_1) is established which, in its most general sense is "Not H_0". That is, the general alternative hypothesis states that, one way or another, at least one of the J population means must differ from one at least one of the others. Mathematically such an alternative hypothesis may be viewed as composite hypothesis, representable in J dimensions corresponding to the values of the J population means.
3. The investigator computes a single "summary score", which constitutes evidence that the null versus the alternative hypothesis is correct. Generally, the greater is the value of the summary score, the greater is the evidence that the alternative hypothesis is true. In the present example—a one-way ANOVA design—the summary score is an F-ratio which is proportional to the variance among the sample means. A small F constitutes evidence for H_0, while the larger is the F, the greater is the evidence for H_1.
4. The sampling distribution of the summary score is determined under the assumption that H_0 is true.
5. A criterion summary score is determined such that, if H_0 is correct, the obtained value of the summary score will be achieved or exceeded with some small probability referred to as α (traditionally, $\alpha = .05$).
6. The obtained value of the summary score is computed from the data.
7. If the obtained summary score equals or exceeds the criterion summary score, a decision is made to reject the null hypothesis, which is equivalent to accepting the alternative hypothesis. If the obtained summary score is less than the criterion summary score, a decision is made to fail to reject the null hypothesis.
8. By this logic, the probability of rejecting the null hypothesis given that the null hypothesis is actually true (thereby making what is known as a "Type-I error") is equal to α. As indicated, α is set by the investigator via the investigator's choice of a suitable criterion summary score. Given that the alternative hypothesis is true, the probability of failing to reject H_0 is known as a Type-II error. The probability of a Type-II error is referred to as β. Closely related to β is $(1-\beta)$ or *power*, which is the probability of correctly rejecting the null hypothesis given that H_1 is true. Typically, β and power cannot be easily measured, because to do so requires a specific alternative hypothesis, which typically is not available[3].

Problems with the LM and with NHST

The LM can be used without proceeding on to NHST and NHST can be used with models other than the LM. However, a conjunction of the LM and NHST is used in the vast majority of experiments within the social sciences and in other sciences, notably the medical sciences, as well. Both the LM and NHST have shortcomings with respect to the insight into a data set that they provide. However, it is my opinion that the shortcomings of NHST are more serious than the shortcomings of the LM. In the next two subsections, I will briefly describe the problems with the LM, and I will then provide a somewhat lengthier discussion of the problems with NHST.

---

with the mean, the overall error variance cannot be assumed to be fully constant. Nonetheless, the linear model formulated would still be a very useful approximation.

[3] More precisely, power can be represented as a function over the J-dimensional space, mentioned earlier, that corresponds to the J-dimensional alternative hypothesis.

Problems with the LM

The LM is what might be termed an off-the-shelf model: That is, the LM is at least a plausible model that probably bears at least some approximation to reality in many situations. However, its pervasiveness often tends to blind investigators to alternative ways of representing the psychological processes that underlie the data in some experiment.

More specifically, although there are different LM equations corresponding to different experimental designs, all of them are *additive with respect to the dependent variable*: that is the dependent variable is assumed to be the sum of a set of theoretical parameters (see, for example, Table 1 and Equation 1, below). The simplicity of this arrangement is elegant, but it de-emphasizes other kinds of equations that might better elucidate the underlying psychological processes.

I will illustrate this point in the context of the classic question: What is the effect of degree of original learning on subsequent forgetting and more particularly, does forgetting *rate* depend on degree of original learning? My goal is to show how the LM leads investigators astray in their attempts to answer this question, and that an alternative to the LM provides considerably more insight.

Slamecka and McElree (1983) reported a series of experiments with the goal of determining the relation between degree of original learning and forgetting rate. In their experiments subjects studied word lists to one of two degrees of proficiency. Subjects' memory performance then was measured following forgetting intervals of 0, 1, or 5 days. Within the context of the LM, the relevant equation relating mean performance, μ_{jk} to delay interval j and initial learning level k is,

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk} \quad \text{(Eq. 1)}$$

where α_j is the effect of delay interval j (presumably α_j monotonically decreases with increasing j), β_k is the effect of degree of learning k (presumably β_k monotonically increases with increasing k) and finally, γ_{jk}, a term applied to each combination of delay interval and learning level, represents the interaction between delay interval and learning level.

Within the context of the LM, two theoretical components are construed as independent if there is no interaction between them. In terms of Equation 1, degree of learning and forgetting are independent if all the γ_{ij}'s are equal to zero. The critical null hypothesis was tested by Slamecka and McElree was therefore that $\gamma_{ij} = 0$ for all i, j. They used their resulting failure to reject this null hypothesis as evidence for the proposition that forgetting rate is independent of degree of original learning.

This conclusion is dubious for a variety of reasons. For present purposes, I want to emphasize that Slamecka and McElree's analysis technique (which Slamecka, 1985, vigorously defended) emerged quite naturally from the LM-based Equation 1 above. Because the LM is so simple, and is so ingrained as a basis for data analysis, it seemed, and still seems, unnatural for workers in the field to consider alternatives to the LM.

What would such an alternative look like? In the final section of this chapter, I will provide some illustrations of alternatives to the LM. In the present context, I will briefly discuss an alternative model within which the learning-forgetting independence issue can be investigated. This model, described by Loftus (1985a; 1985b; see also Loftus & Bamber, 1990) rests on an analogy to forgetting of radioactive decay. Consider two pieces of radioactive material, a large piece (say 9 gms) and a small piece (say 5 gms). Suppose the decay rates are the same in the sense that both can be described by the equation,

$$M = M_0 e^{-kd} \quad \text{(Eq. 2)}$$

where M is the remaining mass after an interval of d days, M_0 is the original mass, and k is the decay constant[4].

The Equation-6 decay curves corresponding to the two different chunks are shown in Figure 1, with the same decay constant, k=0.5, describing the two curves. These curves could, of course, be described by the LM (Equation 1). The γ_{jk} terms would be decidedly nonzero, reflecting the interaction that is represented in Figure 1 by the decreasing vertical distance between the two decay curves with increasing decay time. Thus, using the LM, and Slamecka and McElree's logic, one concludes that large-chunk decay is faster than small-chunk decay.

This conclusion would, in a very powerful sense, be incorrect: As noted above, the Figure-1 decay curves were generated by equations having identical decay rates (k = 0.5). The key to understanding this error is that independence of radioactive decay rates is not associated with lack of interaction within the context of the LM. Instead, it is associated with another kind of lack of in-

[4] This is not a technically correct description of radioactive decay, as radioactive material actually decays to some inert substance instead of to nothing, as implied by Equation 2. For the purposes of this discussion, the "decaying material" may be thought of as that portion of the material that actually does decay, and the logic is unaffected.

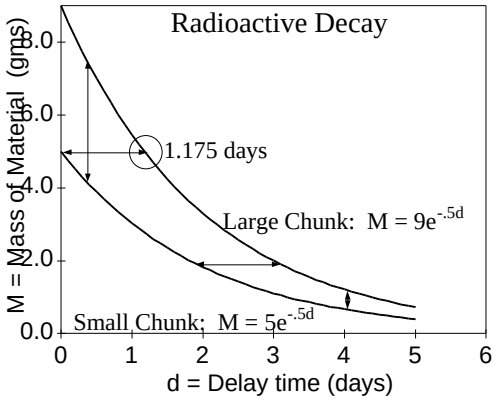

Figure 1. Radioactive decay curves. The decay rate is the same (k = 5) for both the large chunk (9 units) and small chunk (5 units). Note that the vertical distance between the curves decreases over decay time, while the horizontal distance between the two curves is independent of amount of decay time.

teraction which can be intuitively understood as follows. Consider the large chunk. After some time period (which is approximately 1.175 days, as indicated in Figure 1) the large chunk has decayed to the point where only 5 gms remain—i.e., it is physically identical to what the small chunk had been at time zero. Therefore, the large-chunk decay curve following time 1.175 days must be identical to the small-chunk decay curve following time zero; that is, the two decay curves are *horizontally parallel*, separated by a constant delay of 1.175 days. This corresponds to "no interaction" in the horizontal rather than the LM-oriented vertical sense.

The original question, "What is the effect of learning rate on memory" can now be addressed using the same logic, within the context of which forgetting curves resulting from different degrees of original learning must be compared horizontally rather than vertically. The finding of horizontally parallel curves implies that forgetting rate is independent of degree of original learning, while horizontally nonparallel curves imply that forgetting rate depends on degree of original learning[5].

The general model to be tested, given this logic is,

$$\mu(L_1, d_j) = \mu[L_2, f(d_j)] \quad \text{(Eq. 3)}$$

where $\mu(X, d_j)$ refers to mean performance at learning level X following delay interval d_j, and $f(d_j)$ is some function of d_j. Of interest is the nature of the function f on the right side of Equation 3. Various possibilities can be considered. A finding of $f(d_j) = d_j$, would imply no effect at all of original learning on performance. A finding of $f(d_j) = d_j + c$, $c \neq 0$, would imply that forgetting rate is independent of degree of original learning: The curves are parallel, separated by some interval c. Finally, a finding of $f(d_j) = d_j + c + \alpha_j$ where α_j is an amount that varies with d_j, would imply that forgetting rate depends on degree of original learning: The curves are not horizontally parallel.

To summarize, the LM is widely used and probably an approximately correct description of many experimental situations. However it is not always the best model within which an experimental situation can be described, and it is sometimes seriously misleading. It is imperative to realize that one is not bound by the LM just because it is pervasive.

Problems with Null Hypothesis Significance Testing

Upon stepping down as editor of the *Journal of Experimental Psychology*, Arthur Melton published a highly influential editorial (Melton, 1962). In this editorial, Melton emphasized that the criteria used by his journal for accepting manuscripts revolved heavily around NHST, pointing out that (1) articles in which the null hypothesis was not rejected were almost never published and (2) rejection at the .05 significance level was rarely adequate for acceptance; rather, rejection at the .01 level was typically required.

This is a remarkable position. Essentially, it places the process of NHST not only at the heart of data analysis but also the heart of personal scientific advancement: If you don't reject null hypotheses, you don't publish. It is little wonder that NHST is so pervasive in psychology.

Over the past half-century, periodic articles have questioned the value of NHST[6]. Until recently, these articles seem to have had little effect on the means by which data analysis has been car-

[5] For ease of exposition, I have assumed exponential decay in this description. However, as proved by Loftus (1985a, Appendix 2) the implication of independence from horizontally parallel curves does not require the assumption of exponential decay.

[6] A sample of these writings is, in chronological order: Tyler (1931); Jones (1955); Nunnally (1960); Rozebloom (1960); Grant (1962); Bakan (1966); Meehl (1967); Lykken (1968); Carver (1978); Meehl (1978); Berger & Berry (1988); Hunter & Schmidt (1989); Gigerenzer, et al. (1989); Rosnow & Rosenthal (1989); Cohen (1990); Meehl (1990); Loftus (1991; 1993); Carver (1994); Cohen (1994); Loftus & Masson (1994); Maltz (1994); Loftus (1995; 1996); Schmidt (1996); Schmidt & Jones (1997); and Harlow, Mulaik, & Steiger (1997).

ried out. Over the past 10 years, however, there has at least been some recognition of the issues raised by these articles; this recognition has resulted in APA and APS task forces and symposia on the topic, editorials explicitly questioning the use of NHST (e.g., Loftus, 1993b), and occasional calls for the banning of NHST (with which I do not agree), along with a small but still dimly perceptible shift away from exclusive reliance on NHST as a means of interpreting and understanding data.

As I have suggested earlier in this chapter, problems with the LM, such as those described above, pale in comparison to problems with NHST. These problems have been reviewed in the books and articles cited in Footnote 3, and it is not my goal here to provide a detailed rehash of them. Instead, I will sketch them here briefly; the reader is referred to the cited articles for more detailed information. I should note, in the interests of full disclosure, that a number of well reasoned arguments have been made in favor of assigning NHST at least a minor supporting role in the data-comprehension drama. The reader is directed to Abelson (1995) and Krantz (1999) for the best of such arguments.

The major difficulties with NHST are these.

Information Loss as a Result of Binary Decision Processes

A data set is often quite rich. As a typical example, a 3x5 factorial design contains 15 conditions and hence 15 sample means to be accounted for (ignoring of course, per the LM, the raw data from within each condition along with less favored statistics such as the variance, the kurtosis and so on). However, a standard ANOVA reduces this data set to three bits of information: Rejection or failure to reject the null hypotheses corresponding to the effects of Factor 1, Factor 2, and the interaction. Granted, one can carry out additional post-hoc tests or simple-effects tests, but the end result is still that the complex data set is understood, via the NHST process, only in terms of a series of binary decisions rather than as a unified pattern. This is a poor basis for acquiring the kind of gestalt that is necessary for insight and gut-level understanding of a data set.

The Implausibility of the Null Hypothesis

Consider the hypothetical experiment described at the beginning of this chapter. There were five conditions, involving five exposure durations in a free-recall experiment. In a standard ANOVA, the null hypothesis would be:

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \quad \text{(Eq. 4)}$$

where the μ_j's refer to the population means of the five conditions. Note here that "=" signs in Equation 4 must be taken seriously: Equal means *equal* to an infinite number of decimal places. If the null hypothesis is fudged to specify that "the population means are about equal" then the logic of NHST collapses, or at least must be supplemented to include a precise definition of what "about equal" means.

As has been argued by many, a null hypothesis of the sort described by Equation 4 cannot literally be true. Meehl (1967) makes the argument most eloquently, stating,

> Considering...that everything in the brain is connected with everything else, and that there exist several 'general state-variables' (such as arousal, attention, anxiety and the like) which are known to be at least *slightly* influenceable by practically any kind of stimulus input, it is highly unlikely that *any* psychologically discriminable situation which we apply to an experimental subject would exert literally *zero* effect on any aspect of performance." Alternatively, the μ_j's can be viewed as measurable values on the real-number line. Any two of them being identical implies that their difference (also a measurable value on the real-number line) is exactly zero—which has a probability of zero.[7]

And therein lies a serious problem: It is meaningless to reject a null hypothesis that is impossible to begin with. An analogy makes this clear: Suppose an astronomer were to announce that "Given our data, we have rejected the null hypothesis that Saturn is made of green cheese." Although it is unlikely that this conclusion would be challenged, a consensus would doubtless emerge that the astronomer must have been off his rocker for even considering such a null hypothesis to begin with. Strangely, psychologists who make equally meaningless statements on a routine basis

---

[7] A caveat is in order here. Most null hypotheses are of the sort described by Equation 4; that is, they are *quantitative*, specifying a particular set of relation among a set of population parameters. It is possible, in contrast, for a null hypothesis to be *qualitative* (see, e.g., Frick, 1995, for a discussion of this topic). An example of such an hypothesis, described by Greenwald, et al., 1996, is that the defendant in a murder case is actually the murderer. This null hypothesis could certainly be true; however; the kind of qualitative null hypothesis that it illustrates constitutes the exception rather than the rule.

continue to be regarded as entirely sane. (Even stranger is the common belief that an α-level of .05 implies that an error is made in 5% of all experiments in which the null hypothesis is rejected. This is analogous to saying that, of all planets reported not to be made of green cheese, 5% of them actually *are* made of green cheese.)

Decision Asymmetry

Putting aside for the moment the usual impossibility of the null hypothesis, there is a decided imbalance between the two types of errors that can be made in a hypothesis-testing situation. The probability of a Type-I error, α, can be, and is, set by appropriate selection of a summary-score criterion. However, the probability of a Type-II error, β, is, as noted earlier, generally unknowable because of the lack of a quantitative alternative hypothesis. The consequence of this situation is that rejecting the null hypothesis is a "real" decision, while failing to reject the null hypothesis is, as the phrase suggests, a nondecision: It is simply an admission that the data do not provide sufficient information to support a clear decision.

Accepting H_0

The teaching of statistics generally emphasizes that "we fail to reject the null hypothesis" does not mean the same thing as "we accept the null hypothesis". Nonetheless, the temptation to accept the null hypothesis (usually implicitly so as to not brazenly disobey the rules) often seems to be overwhelming, particularly when an investigator has an investment in such acceptance. As I have noted in the previous section, accepting a typical null hypothesis involves faulty reasoning anyway because a typical null hypothesis is impossible. However, particularly in practically-oriented situations, an investigator is justified in accepting the null hypothesis "for all intents and purposes" assuming that the investigator has convincingly shown that there is adequate statistical power (see Cohen, 1990; 1994). Such a power analysis is most easily carried out by computing some kind of confidence interval (described in detail below) which would allow a meaningful conclusion such as "the population mean difference between Conditions 1 and 2 is, with 95% confidence, between $\pm\varepsilon$" where ε is a sufficiently small number that the actual difference between Conditions 1 and 2 is inconsequential from a practical perspective.

The misleading dichotomization of "$p < .05$" vs. "$p > .05$" results

As indicated in his 1962 editorial, summarized earlier, Arthur Melton considered an observed p-value of .05 to be maximal for acceptance of an article. Almost four decades later, more or less this same convention holds sway: Who among us has not observed the heartrending spectacle of a student or colleague struggling to somehow transform a vexing 0.051 into an acceptable 0.050?

This is bizarre. The actual difference between a data set that produces a p-value of 0.051 versus one that produces a p-value of 0.050 is, of course, miniscule. Logically, very similar conclusions should issue from both data sets, and yet they do not: The .050 data set produces a "reject the null hypothesis" conclusion, while the .051 data set produces a "fail to reject the null hypothesis" conclusion. This is akin to a chaotic situation in which small initial differences distinguishing two situations lead to vast and unpredictable eventual differences between the situations.

The most obvious consequence of this situation is that the lucky recipient of the .050 data set gets to publish, while his unlucky .051 colleague does not. There is another consequence, however, which is more subtle but probably more insidious: The reject/fail-to-reject dichotomy keeps the field awash in confusion and artificial controversy. This is because investigators, like most humans, are loath to make and stick to conclusions that are both weak and complicated, like "we fail to reject the null hypothesis." Instead investigators are prone to (often unwittingly) transform the conclusion into the stronger and simpler, "we accept the null hypothesis." Thus two similar experiments—one in which the null hypothesis is rejected and one in which the null hypothesis is not rejected—can and often do lead to seemingly contradictory conclusions—"the null hypothesis is true" versus "the null hypothesis is false." The inevitable head-scratching and subsequent flood of "critical experiments" that are generated by such "failures to replicate" may well constitute the single largest source of wasted time in the practice of psychology.

The counternull

Robert Rosenthal (see chapter this volume) has suggested a simple score called the "counternull" which serves to underscore the difficulty in accepting H_0. The counternull revolves around an increasingly common measure called "effect size," which, essentially is the mean magnitude of some effect (e.g., the mean difference between two conditions) divided by the standard deviation (generally pooled over the conditions). Obviously, all else equal, the smaller the effect size, the less inclined one is to reject H_0. Suppose, to illustrate, that in some experiment one found an effect size of 0.20 which was insufficiently large to reject H_0. As noted earlier, the temptation is often overwhelming to accept H_0 in such a situation because the world seems so much clearer that way. It is therefore useful to report Rosenthal's counternull

which is simply twice the effect size or 0.40 in this example. It is sobering to realize that the data permit a reality corresponding to the counternull (0.40) just as much as they permit a reality corresponding to H_0 (an effect size of zero). The use of the counternull also subtly underscores a fact which is almost invisible in a NHST framework, specifically that the best estimate of some population parameter is the corresponding statistic that is measured in the experiment. So in this example, the best estimate of the population effect size is exactly what was measured—0.20—rather than the zero value toward which the investigator is drawn in an hypothesis-testing framework.

The p(data|H_0) versus p(H_0|data) Confusion

In the previous section, I discussed the critical consequences of having a data set that produces p = .050 versus one that produces p = .051. What exactly is it that these p values refer to?

To address this question let us again set aside the awkward fact of the null hypothesis's usual impossibility and suppose that the null hypothesis actually has a reasonable possibility of being true. It is taught in every statistics class that a p-value less than .05 means that,

$$p = p(data|H_0) < .05 \qquad \text{(Eq. 5)}$$

So what do you do with a sufficiently small p-value? You reject the null hypothesis. What does it mean to reject the null hypothesis? In everyday language, to reject the null hypothesis in light of the data means pretty unequivocally that given the data, the probability of the null hypothesis is so low that it should be rejected, i.e., that

$$p(H_0|data) \text{ is small} \qquad \text{(Eq. 6)}$$

Thus it should come as no surprise that the sacred .05 is often incorrectly associated with the conditional probability of Equation 6 rather than correctly associated with the opposite conditional probability of Equation 5.

Now indeed, if $p(data|H_0) < .05$, then it is likely true that $p(H_0|data)$ is also smallish: After all, since

$$p(H_0|data) = \frac{p(H_0 \cap data)}{p(data)}$$

and

$$p(data|H_0) = \frac{p(H_0 \cap data)}{p(H_0)}$$

the two conditional probabilities share the same numerator and are therefore somewhat related to one another. However, the probability that the investigator is primarily interested in, $p(H_0|data)$, is not known to any degree of precision. It is therefore breathtakingly silly to place such vast emphasis on the exact value of $p(data|H_0)$ when this probability is only indirectly interesting to begin with.

SUGGESTED DATA ANALYSIS TECHNIQUES

I now turn to a description of six data-analysis techniques which are considerably more useful than strict adherence to NHST in their ability to illuminate a data set's meaning and to answer whatever question originally prompted the experiment. The first two of these—the use of pictorial representations and use of confidence intervals—are not novel; they are just not widely used, or at least are not widely used to best advantage. The third and fourth techniques—use of planned comparisons and other means of accounting for different sources of variance—are also not novel, but are hardly ever used. The fifth—use of mathematical process models—has an honorable tradition in the area of mathematical psychology, but is still not pervasive. The final set of techniques, which I have termed equivalence techniques, are standard in vision science, but are almost never used in other areas of psychology.

Pictorial Representations

If the results of an experiment consist of more than two numbers, then providing some form of pictorial representation of them is enormously useful in providing a reader with an overall, gestalt image of what the data are all about. (This seems so obvious that it seems hardly worth saying, but the obviousness of the concept does not always translate into the concomitantly obvious behavior.)

To illustrate, Table 2 and Figure 2 show the same data set (response probabilities from a hypothetical experiment in which digit strings are presented for varying durations and contrasts) as a table and as a figure. It is obvious that the table

Table 2. Data (proportion correct) for an experiment in which stimuli are presented at one of six durations and one of three contrast levels.

	Contrast		
Duration (ms)	0.05	0.10	0.20
10	0.069	0.134	0.250
20	0.081	0.267	0.375
40	0.230	0.466	0.741
80	0.324	0.610	0.872
160	0.481	0.768	0.898
320	0.574	0.799	0.900

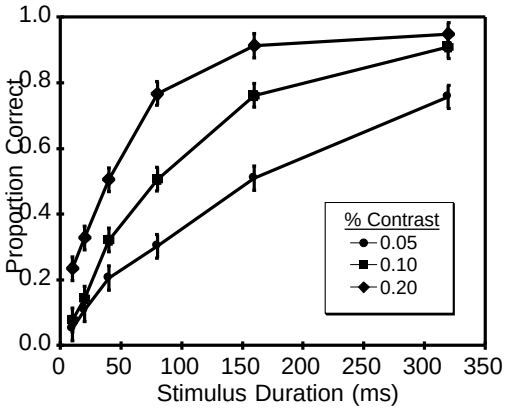

Figure 2. Hypothetical data from a 5 (stimulus exposure duration) x 3 (stimulus contrast level) experiment. The dependent variable is proportion correct recall. Error bars represent standard errors.

can only be understood (and not very well understood at that) via a lengthy serial inspection of the numbers within it. In contrast a mere glance at the corresponding figure renders it entirely clear what is going on.

Graphs Versus Tables

Despite the obvious and dominating expositional advantage of figures over tables, data continue to be presented as tables are at least as often, or possibly more often than as figures. For most of the psychology's history, the reason for this curious practice appeared to be founded in a prosaic matter of convenience: While it was relatively easy to construct a table of numbers on a typewriter, constructing a decent figure was a laborious undertaking. You drew a draft of the figure on graph paper, took the draft to an artist who invariably seemed to reside on the other side of the campus, following which you waited a week for the artist to produce a semi-finished version. Then you made whatever changes in the artist's rendering that seemed appropriate. Then, you repeatedly iterated through this dreary process until the figure was eventually satisfactory. Finally, adding insult to injury, you had to take the finished drawing somewhere else to have its picture taken before the publisher would take it. Who needed that kind of hassle?

Today, obviously, things are much different, as electronic means of producing figures abound. To obtain information about popular graphing techniques, I conducted an informal survey in which I emailed to all researchers in my email address book, a request that they tell me what graphing technique(s) they use. One hundred and sixty one respondents used a total of 229 techniques, and the summarized results are provided in Table 3.

Table 3. Techniques for plotting data, as revealed by an informal survey.

Application Name	Frequency
Microsoft Excel	55
CricketGraph	27
SigmaPlot	22
KaleidaGraph	17
SPSS	16
MATLAB	15
PowerPoint	10
DeltaGraph	9
S-plus	7
Mathematica	5
Microsoft Office	5
Systat	5
Igor/Igor Pro	4
Statistica	4
Gnuplot	3
Canvas	2
Hand plotting	2
StatView	3
ABC Graphics	1
Autocad	1
Axum	1
c graph-pac	1
ClarisDraw	1
Grapher	1
Graphpad	1
Illustrator	1
JMP	1
MacDraw	1
Maple 2D	1
Origin	1
PsiPlot	1
Quattro Pro	1
R	1
SciPlot	1
Smartdraw	1
TK solver	1

The results of this survey can be summarized as follows. Fewer than 25% of the application programs mentioned were statistical packages, perhaps because the most commonly used packages do not provide very flexible graphing options. Over a third of the applications were specialized drawing programs (CricketGraph, SigmaPlot, and KaleidaGraph were the most popular, but many others were mentioned). About 10% of the applications were general-purpose presentation programs (Powerpoint was the most popular) and the final one-third was general-purpose analysis programs, with Microsoft Excel accounting for

the majority of these instances. Excel was by far the single leading application used for graphing. Seven respondents reported never graphing data, while 13 assigned the task to someone else. Two people reported still drawing graphs by hand. The remaining 139 respondents used some form of electronic graphing techniques.

At the present time, a brief description of graphing programs is supplied by Denis Pelli (personal communication) and can be found at http://vision.nyu.edu/Tips/RecSoftware.html.

Graph-Making Transgressions

I have tried to present a fairly bright picture of the ease of creating high-quality graphs. There is, however, a dark side of this process which is that a graph-creator has the capability of going wild with graphical features, thereby producing a graph that is difficult or impossible to interpret. For example David Krantz (personal communication) has noted that, for example, graphmakers often attempt to pack too much information into a graph, they produce graphs that are difficult to interpret without intense serial processing, they produce unintended and distracting emergent perceptual features, or they simply omit key information either in the graph itself of in the graph's legend. There are, of course, many other such transgressions, treatments of which are found in the references provided in the next section. (My own personal bête noire is the 3-D bar graph.)

Other Graphical Representations

A discussion of graphs is limited in the sense that there are myriad means of visually presenting the results of a data set. It is beyond the scope of this chapter to describe all of them. For an a set of initial pointers to a set of sophisticated and elegant graphical procedures, the reader is directed to excellent discussions and examples in Tufte (1983; 1990), Tukey (1977), and Wainer & Thissen (1993). The main point I want to make is that pictorial representations almost always excel over their verbal counterparts as an efficient way of conveying the meaning of a data set.

The Use of Confidence Intervals

Earlier, I described the LM as the standard model for linking a data set to the answer to the scientific question at hand. Somewhere in a LM equation (e.g., Equation 1) are always one or more error terms which represent the uncertainty in the world.

Using the LM to answer scientific questions is a two-stage process. The first stage is to somehow determine knowledge of relevant population parameters given measured sample statistics along with the inevitable statistical noise. The second stage is to use whatever knowledge emerges about population parameters to answer the question at hand as best as possible.

It seems almost self evident that that the second stage—deciding the implications of the pattern of population parameters for the answer to the question at hand—should be the investigator's fundamental goal. In contrast, the typical routine of statistical analysis—carrying out some procedure designed to cope with the noise-limited relation between the sample statistics and the corresponding population parameters—should be viewed as a necessary but boring nuisance. If the real world suddenly transformed into an ideal world in which experiments produced no statistical noise, it would be cause for rejoicing among investigators, as a major barrier to data interpretation would be absent.

There are two basic procedures for coping with statistical noise in quest of determining the relations between a set of sample statistics and their population counterparts. The first procedure entails attempting to determine what the pattern of population parameters *isn't*, i.e., trying to reject a null hypothesis of some specific, usually uninteresting, pattern of population parameters, via NHST. The second procedure entails attempting to determine what the pattern of population parameters *is*, using the pattern of sample statistics as an estimate of the corresponding pattern of population parameters, along with error bars to represent the degree of conclusion-obscuring statistical noise. It is my (strong) opinion that trying to determine what something is is generally more illuminating than trying to determine what it isn't.

The use of error bars, e.g., in the form of 95% confidence intervals, around plotted sample statistics (usually sample means) is an ideal way of presenting data in such a way that the results of both these two data-analysis and interpretation stages are represented and that their relative importance is depicted. Consider a plot such as the one shown in Figure 2. The pattern of sample means represents the best estimate of the corresponding pattern of population means. This pattern is fundamental to understanding how perception is influenced by contrast and duration and it is this pattern that is most obvious and fundamental in the graph. Secondarily, the confidence intervals provide a quantitative visual representation of the faith that should be placed in the pattern of sample means as an estimate of the corresponding pattern of population means. Smaller confidence intervals, of course, mean a better estimate: In the extreme, if the confidence intervals were of zero length, it would be clear that error was irrelevant, and that the investigator could spend all his or her energy on the fundamental task of figuring out the impli-

cations of the pattern of population means for answering the questions at hand.

The Interpretation of a Confidence Interval

The technically correct interpretation of a confidence interval is this. Suppose that many random samples of size n are drawn from some population. For each sample the sample mean, M is computed, and a confidence interval—suppose, for simplicity of exposition, a 95% confidence interval—is drawn around each mean. Approximately 95% of these confidence intervals will include μ, the population mean.

Returning now to Planet Earth, what does this logic imply in the typical case wherein a single mean is computed from a single sample, and a single confidence interval is plotted around that sample mean? If the confidence interval were the only information available to the investigator, then the investigator would conclude that, with 95% probability, this confidence interval is one of the 95% of all possible confidence intervals that include μ; i.e., the simple conclusion can be drawn that with 95% probability the confidence interval includes μ.

However, the caveat must be issued that sometimes an investigator *does* have additional information available (such information is, for instance, the basis for doing a one-tailed rather than a two-tailed test). In this case, the investigator's subjective probability that confidence interval contains a population parameter may be influenced by this additional information as well as by the confidence interval itself. For instance, an investigator examining a 95% confidence interval constructed around a particular sample mean may, based on such other information, be skeptical that it does in fact contains μ. Whether or not an investigator chooses to quantify such beliefs using probabilities, it is sometimes misleading to state unequivocally after examining the data, that the particular interval has a 95% probability of including μ.

Despite this caveat, however, construal of an x% confidence interval as including the population parameter with x% probability is generally a reasonable rule of thumb (as distinguished from something like, "since $p < .05$, H_0 is likewise true with a probability of less than about .05," which is definitely *not* a reasonable rule of thumb).

Confidence Intervals Around Linear Combinations of Variables

For many of the examples to follow, it is important to remind the reader of the relation between a confidence interval around a single mean, and a confidence interval around a linear combination of means. In particular, suppose an experiment results in a series of means, M_1, M_2,…, M_J. If the confidence interval around any of the M_j's has a length of X, then the confidence interval around any linear combination of the means, $k_1M_1 + k_2M_2$,…, $+ k_JM_J$, has a length of,

$$X\sqrt{k_1^2 + k_2^2 + ... + k_J^2} \qquad \text{(Eq. 7)}$$

The most frequent use of the property described by Equation 7 is when a confidence interval around a difference score, (M_1-M_2) is desired. In this situation, $k_1 = 1$, $k_2 = -1$, and the difference-score confidence interval is therefore the individual-mean confidence interval multiplied by $\sqrt{2}$. Some additional implications of this fact will be provided later in this chapter.

Confidence Intervals and Statistical Power

Within the context of NHST, the definition of power is simple: As indicated earlier, it is the probability of correctly rejecting the null hypothesis given that the null hypothesis is false. However (despite frequent requests on the part of journal editors) explicit power analyses rarely make their way into journal pages. The reasons for this deficit appear to be twofold. First, to compute an exact value of power requires a quantitative alternative hypothesis which is almost never available. Second, the concept of power, while seemingly straightforward is, as anyone who has tried to teach it well knows, almost impossible to get across to anyone who hasn't somehow figured it out already. Many educators and authors give up on the topic; for instance, Guilford (1942) in his widely read *Fundamental Statistics in Psychology and Education* declared power to be "too complicated to discuss."

As has been frequently noted, the issue of power is particularly important if a scientific conclusion entails the acceptance of some null hypothesis. In such a situation, it is incumbent on the investigator to convince his or her audience that the power of the relevant statistical test is high. How is this to be done?

Because there is indeed a profound dearth of quantitative alternative hypotheses in the social sciences, a single value of power typically cannot be computed. Therefore, some more general representation of power must be concocted for a particular experiment. One occasionally suggested such representation involves the use of power curves (e.g., Hays, 1973; p. 359) whereby power is plotted as a function of the value of the alternative hypothesis.

Another way of representing power is via the magnitude of confidence intervals. The rule here is simple: greater the statistical power, the smaller are the confidence intervals. To illustrate, imagine

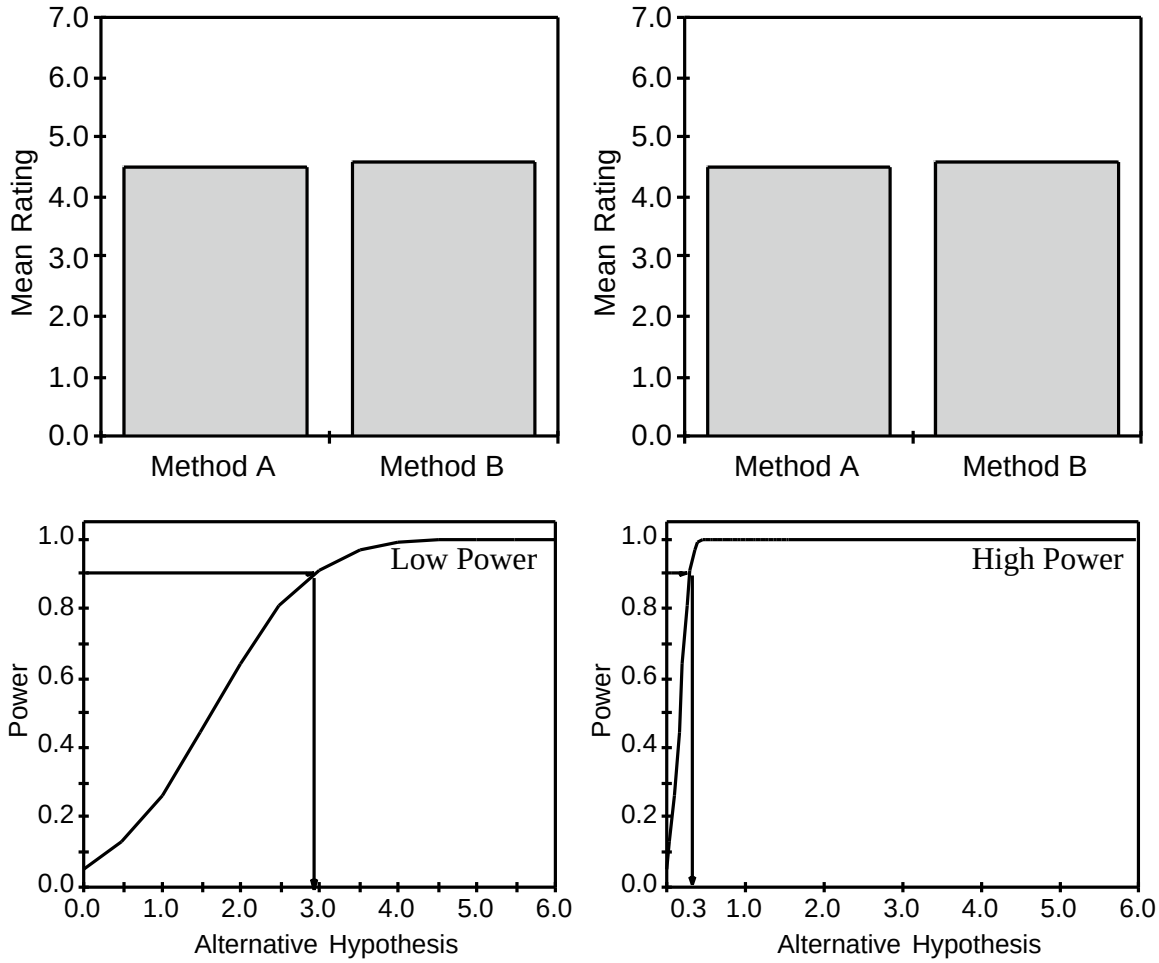

Figure 3. One technique for carrying out a power analysis. A low-power situation is in the left panels) and a high-power situation is in the right panels. The top panels show the data, the while the bottom panels show power curves.

a hypothetical experiment in which a clinical researcher is investigating the relative effectiveness of two methods, Method A and Method B, of reducing anxiety. Two groups of high-anxiety subjects participate in the experiment, one receiving Method A and the other receiving Method B. Following their treatment, subjects rate their anxiety on a 7-point scale. Suppose that the experiment results in a small, not statistically significant difference between the two methods. In what follows, I will demonstrate two techniques of presenting the results for two hypothetical cases: A low-power case involving n subjects, and a high-power case involving 100n subjects.

The first analysis technique incorporates standard NHST, along with a formal power analysis. Figure 3 shows the graphical results of this kind of analysis for the low-power case (left panels) and the high-power case (right panels). The top panels show bar graphs depicting the main experimental results while the bottom panels show power curves that depict power as a function of the difference between two population means according to a continuous succession of alternative hypotheses. Power is represented by the slope of the power curves. As illustrated by the arrows, the low-power curve achieves a power of 0.90 when the alternative hypothesis is that the population means differ by about 3.0, while the high-power curve achieves 0.90 when the alternative hypothesis is that the population means differ by about 0.3.

Figure 4 shows a different way of representing this power information for the same low-power case (left panel) and high-power case (right panel). Figure 4 again shows the bar graph, but here, the bars are accompanied by 95% confidence intervals around the means that they depict. The free-floating error bars show the magnitude of the 95% confidence interval around the population mean differences in each of the panels. Here, power is

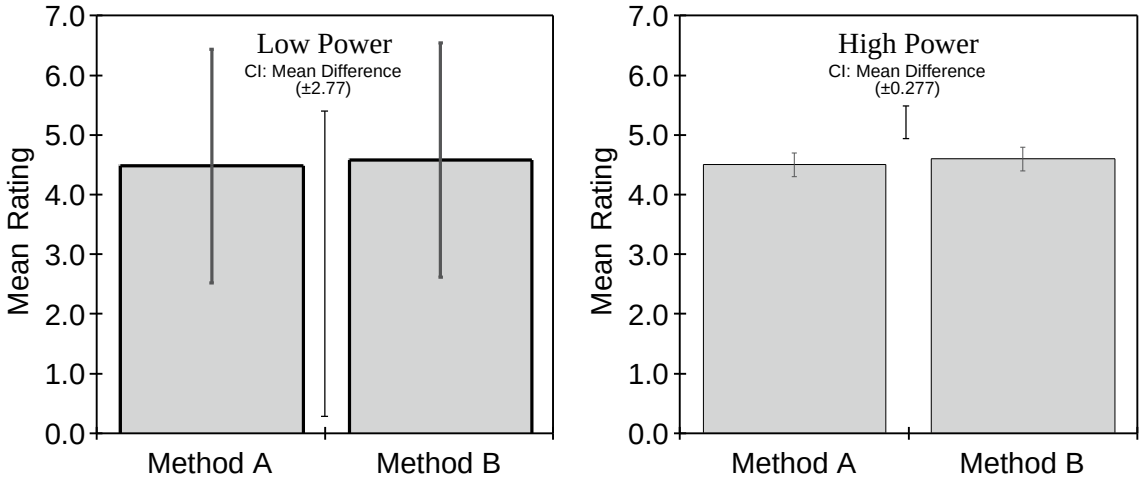

Figure 4. A second technique for carrying out a power analysis in the anxiety treatment method experiment. Smaller confidence intervals reflect greater power.

represented quite simply by the size of the confidence intervals which are large in the left (low-power) graph, but small in the right (high-power) graph.

In short, Figures 3 and 4 show the same information, However, Figure 4 presents the information in a much simpler and more intuitive manner than does Figure 3. Figure 4 makes it immediately and visually clear how seriously the sample means and the sample mean differences are to be taken as estimates of the corresponding population means which, in turn, provides critical information about how "nonsignificance" should be treated. The left panel of Figure 4 leaves no doubt that failure to reject the null hypothesis is a non-conclusion—that there is not sufficient statistical power to make any conclusions at all about the relative magnitudes of the two population means. The right panel, in contrast, makes it evident that something very close to the null hypothesis is actually true—that the true difference between the population means is, with 95% confidence, restricted to a range of 0.277 which is very small in the grand scheme of things.

Confidence Intervals or Standard Errors?

Thus far I have been using 95% confidence intervals in my examples. This is one of the two standard configurations for error bars, the other being a standard error which is approximately a 67% confidence interval[8]. In the interests of standardization, one of these configurations or the other should be used unless there is some compelling reason for some other configuration.

I suggest, in particular, being visually conservative, which means deliberately stacking the deck against concluding whatever one wishes to conclude. This means, one should use 95% confidence intervals, which have a greater effect of suggesting no difference, when the interest is in rejecting some null hypothesis. Conversely, one should use standard errors, which have a greater effect of suggesting a difference, when the interest is in confirming some null hypothesis (as in, for example, when comparing observed to predicted data points in a model fit).

Different Kinds of Confidence Intervals

The interpretation of a confidence interval is somewhat different depending on whether it is used in a between-subjects or a single-factor within-subjects (i.e., repeated-measures) design, a multifactor within-subjects design, or a mixed design (some factors between, other factors within). These differences are discussed in detail by Loftus & Masson (1994). The general idea is as follows.

Between-subjects designs

A confidence interval is designed to isolate a population parameter, most typically a population mean, to within a particular range. A between-subjects design constitutes the usual venue in which a confidence interval has been used in psychology, to the extent that confidence intervals have been used at all. Consider as an example a simple one-way ANOVA experiment in which the investigator is interested in the effects of caffeine on reaction time (RT). Four conditions are defined by four levels of caffeine: 0, 1, 2, or 3 caffeine units per unit body weight. Suppose that n = 10

[8] The exact coverage of a standard error depends, of course, on the number of degrees of freedom going into the error term.

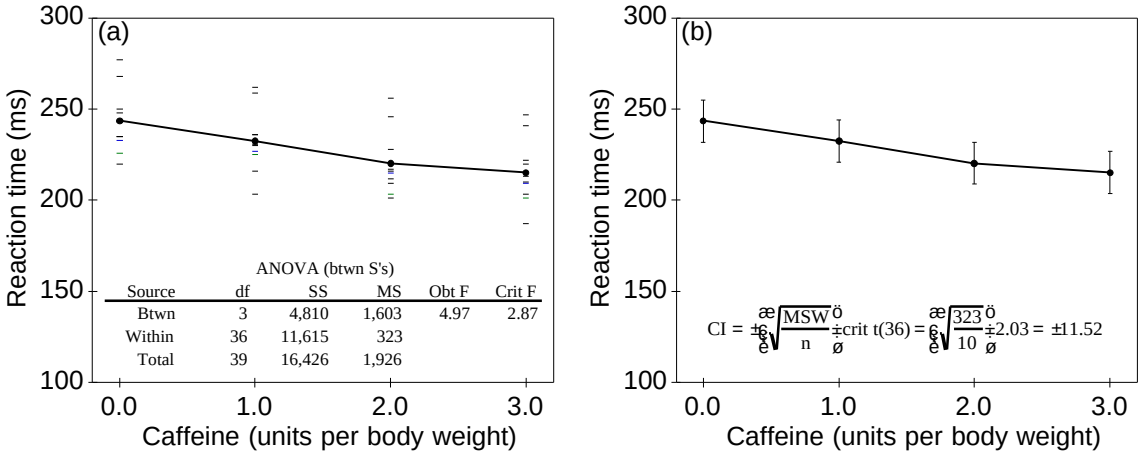

Figure 5. Data from a hypothetical experiment in which RT is measured as a function of caffeine consumption in a between-subjects design. The right panel shows the mean data with each mean surrounded by individual-subject data points. The right panel shows 95% confidence intervals around the sample means.

subjects are randomly assigned to, and participate in each of the four conditions. The outcome of the experiment is as represented in Figure 5a, which shows the mean data (solid line) along with dashes surrounding each mean which represent the 10 individual data points within each condition. The results of an ANOVA are shown at the bottom left of the panel and are straightforward. Note that the total sum of squares which, of course, reflects the total variability of all 40 scores in the experiment is 16,426, and the 39 total degrees of freedom are divided into 3 (between, i.e., caffeine) and 36 (error, i.e., within). (These factoids will become relevant in the next section.)

Computing a confidence interval in such a design is entirely straightforward and is obtained by the equation at the bottom of Figure 5b,

$$CI = \pm\sqrt{\frac{MS(Within)}{n}} \times crit\ t(dfW) \qquad \text{(Eq. 8)}$$

using the MS (Within) from the Figure-5a ANOVA table. The error term going into the confidence interval is the same as in the ANOVA—MS (Within)—and the criterion t is based on dfW, which is 36 in this example. The resulting 95% confidence interval is ±11.52.

Single-factor within-subjects designs

I will now treat the exact same data that I have just described as having come from a within-subject design. That is I will treat the data assuming that each of a total of n = 10 subjects had participated, at one time or another, in each of the 4 conditions. It is now possible to draw an curve relating RT to caffeine for each of the 10 subjects. These curves, along with the same mean curve from Figure 5, are shown in Figure 6a. At the bottom of Figure 6a is the within-subjects ANOVA. Note that the 16,426 total sum of squares (now referred to as "between cells") is divided into caffeine conditions (as with the between-subjects design, equal to 4,810, and based on 3 degrees of freedom), subjects (based on 9 degrees of freedom) and the subject x caffeine interaction (based on 27 degrees of freedom). The relative consistency of the caffeine effect across the different subjects is represented graphically by the relatively parallelness of the individual subject curves, and is represented within the ANOVA by the relatively small interaction (i.e., error) term of MS (Interaction) = 20. The F ratio of 79.72 is considerably greater in this design than it was in the between-subjects design (where F = 4.97). The reason for this is that a large portion of the error variance—the between-subjects variability reflected by SS (Subjects) = 11,072—is irrelevant in this within-subjects design whereas in the between-subjects design, this very same variability formed part of the error term, i.e., was part of SS (Within).

How should a confidence interval be constructed in this kind of within-subjects design? Technically, as described earlier, a confidence interval is designed to isolate a population mean with some degree of probability. In this within-subjects design, the uncertainty of any condition population mean is based on the exactly the same uncertainty as it was in the between-subjects design. More specifically, in the between-subjects design this uncertainty was referred to as "within-condition variance" and in that example, it was SS (Within), based on 36 degrees of freedom. In this within-subjects design, the location of a condition mean is uncertain because of both variability due to subjects, SS (Subjects) = 11,072 based on 9 de-

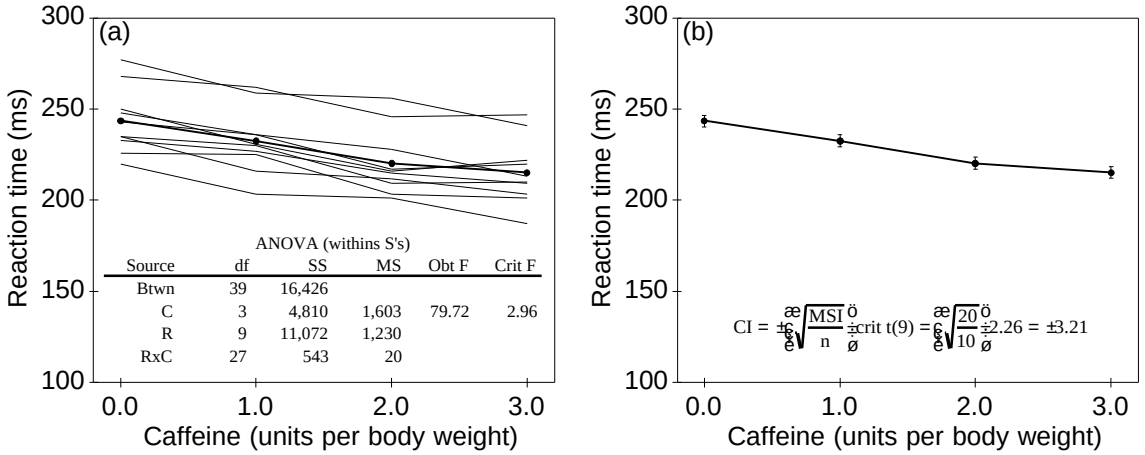

Figure 6. Data from a hypothetical experiment in which RT is measured as a function of caffeine consumption in a within-subjects design. All 40 data points are the same as those shown in Figure 5. The right panel shows the mean data (heavy line) along with individual-subject data points (light lines). The right panel shows 95% "within-subject" confidence intervals around the sample means that is based on the subject x interaction variance.

grees of freedom, and variability due to the subject by condition interaction, SS (Interaction) = 543, based on the remaining 27 degrees of freedom. The combined error variance SS (Subjects plus Interaction) is therefore 11,615, based on 36 degrees of freedom, just as it was in the between-subjects design, and the confidence interval of 11.52 is therefore identical also.

Intuitively this seems wrong. Just as the within-subjects design includes a great deal more sensitivity, as reflected in the substantially greater F ratio in the ANOVA, so it seems that the greater sensitivity should also be reflected in a smaller confidence interval. What is going on?

To answer this question, it is necessary to consider not what a confidence interval is technically, but what a confidence interval is generally used to accomplish. An investigator is not usually interested in absolute values of population means, but rather is interested in *patterns* of population means. So for instance, in the Figures 5 and 6 data the mean RT declines from approximately 240 ms to 215 ms across the caffeine conditions. However, it is not the exact means that are important for determining caffeine's effect on RT; but rather it is the decrease, or perhaps the form of mathematical function describing the decrease that is of interest[9].

This observation has an important implication for the interpretation of confidence intervals: Confidence intervals are rarely used in their "official" role of isolating population means. Instead, they are generally used as a visual aid to judge the reliability of a *pattern* of sample means as an estimate of the corresponding pattern of population means. In the Figure-5 between-subjects data, for instance, the confidence intervals indicate that a hypothesis of monotonically decreasing population-mean RT's with increased caffeine is reasonable.

How does this logic relate to within-subjects designs? The answer, detailed by Loftus and Masson (1994) is that a "confidence interval" based on the interaction variance is appropriate for the goal of judging the reliability of a pattern of sample means as an estimate of the corresponding pattern of population means; thus the within-subjects confidence interval equation is,

$$CI = \pm\sqrt{\frac{MS(Interaction)}{n}}\,crit\ t(dfI) \qquad \text{(Eq. 9)}$$

where n again represents the number of observations on which each mean is based (n = 10 in this example). Using Equation 9 (see Figure 6b), the confidence intervals in Figure 6a were computed using the MS (Interaction) shown in the ANOVA table within Figure 6a. The resulting confidence interval is ±3.21. This value is, of course, considerably smaller than the between-subjects, Figure-5 counterpart of 11.52. It bears emphasis, however, this apparent increase in power comes about because information is lost: In particular the confidence intervals no longer isolate absolute values of population means; rather they are appropriate only

[9] I should note that this is not always true. Sometimes an investigator *is* interested in isolating some population mean. An obvious example would be when the investigator wishes to determine whether performance in some condition is at a chance value.

Table 4. ANOVA table for a two-factor, within-subjects design.

Source	Degrees of freedom	Error term
Factor A (A)	df(A)	MS (A x S)
Factor B (B)	df(B)	MS(B x S)
Inter. (AxB)	df(A x B)	MS(A x B x S)
Subjects (S)	df(S)	
A x S	df(A) x df(S)	
B x S	df(B) x df(S)	
(A x B) x S	df(A) x df(B) x df(S)	

for assessing the reliability of the pattern of sample means as an estimate of the underlying pattern of population means. That is, they serve the same function as they do in the between-subjects ANOVA.

Confidence intervals in multifactor within-subjects designs

In a pure between-subjects design, there is only one error term, MS (Within), irrespective of the number of factors in the design. Therefore, assuming homogeneity of variance, a single confidence interval, computed by Equation 8 or Equation 9, is always appropriate.

In a multifactor within-subjects design, the situation is more complicated in that there are multiple error terms, corresponding to multiple subject-by-something interactions. For instance, in a two-factor within-subjects design, there are three error terms: One corresponding to Factor A, one corresponding to Factor B, and one corresponding to the (AxB) interaction. These error terms are summarized in Table 4, for a standard two-factor, within-subjects design[10]. This raises the problem of how to compute confidence intervals, as it would appear that there are as many possible confidence intervals as there are error terms. Which confidence interval(s) are appropriate to display?

Often the answer to this question is simple, because in many such two-factor designs—and in many multifactor within-subjects designs in general—the error terms are all roughly equal (i.e., differ by no more than a factor of around 2:1). In such instances, it is reasonable to simply pool error terms, that is to compute an overall error term by dividing the sum of the sum of squares (error) by the total degrees of freedom (error) to arrive at a single "subject x condition" interaction, where a "condition" is construed as single combination of the various factors (e.g., a 5 x 3 x subjects design would have 15 separate conditions). This single error term can then be entered into Equation 9 to compute a single interaction. Here "dfI" refers to degrees of freedom in the total interaction between subjects and conditions. So, for instance, in a 5 (Factor A) x 3 (Factor B) x 20 (subjects) design, dfI would be (15-1) x (20-1) = 266. As before, "n" in Equation 9 refers to the number of observations on which each mean is based: 20 in this example.

Of course, Nature is not always this kind, and the investigator sometimes finds that the various error terms have widely varying values. In this situation, the investigator is in a position of having to provide a more complex representation of confidence intervals, and the situation becomes akin to that described in the next section where a mixed design is used.

Confidence intervals in mixed designs

A mixed design is one in which some of the factors are between subjects and other factors are within subjects. For simplicity, I will describe the simplest such design: a two-factor design with one between-subjects factor and one within-subjects factor (see also, Loftus & Masson, 1994, pp. 484-486).

Imagine the caffeine experiment described above except that two different subject populations are investigated: young adults (in their 20's) and older adults (in their 70's); thus there are two variables, one of which (caffeine) is within-subjects and the other of which (age) is between subjects. Again, there are n = 10 subjects in each of the two age groups. Suppose that the data are as depicted in Figure 7a (note that again, the relevant ANOVA table is provided at the bottom of the figure).

As described many standard statistics textbooks, there are two error terms in this design. The error term for the age effect is MS (Subjects within age groups) = 1,656, while the error term for caffeine and for the caffeine x age interaction is the MS (Caffeine x Subjects) = 99. There are, correspondingly, two separate confidence intervals that can be computed. The first, computed by Equation 9, is the kind of "within-subjects" confidence interval that was described in the previous section. This confidence interval which, as indicated at the bottom of Figure 7b is computed to be ±6.3, is appropriate for assessing the observed effects of caffeine and of the age x caffeine interaction as estimates of the corresponding population effects. This confidence interval is plotted around each of the cell means in Figure 7b. Note that this confidence interval is not appropriate for describ-

[10] With more than two factors, the same general arguments to presented below hold, they are simply more complex, because there are yet more error terms; e.g., in a three-factor, within-subjects design, there are 3 main-effect error terms, 3 two-way interaction error terms, and 1 three-way interaction error term, or 7 error terms in all.

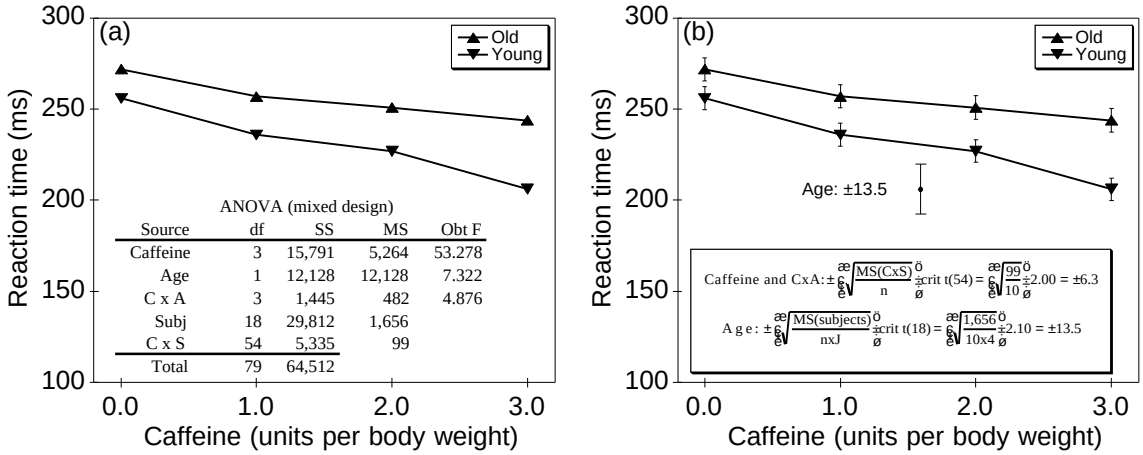

Figure 7. Data from a hypothetical experiment in which RT is measured as a function of caffeine consumption. Caffeine consumption is varied within subjects, and two different age groups are included. The right panel shows 95% "within-subject" confidence intervals around the sample means that is based on the subject x interaction variance, along with a free-floating confidence interval that is appropriate for comparing the two age curves.

ing the absolute age effect. The easiest way to conceptualize what this means is to think of an extreme situation in which the within-subjects confidence interval were zero; thus one could be entirely confident of the nature of the caffeine effect and of the interaction (that is, one could be entirely confident of the shape of each of the two curves in Figure 7). However, the vertical relation of the two age curves to one another would still be uncertain.

How uncertain? This would be determined by the size of the other, "between-subjects" confidence interval, based on MS (Subjects). As shown at the bottom of Figure 7b, the equation for computing this confidence interval is,

$$CI = \sqrt{\frac{MS(Subjects)}{nxJ}}\ \text{crit t(dfS)}$$

or 13.5. The value of this confidence interval, along with a pictorial representation of its value is shown in the middle of Figure 7b. Because there are only two age levels, an alternative way of depicting the age effect would be in terms of the age difference: The confidence interval around a difference score is always equal to the confidence interval around the individual component times $\sqrt{2}$. In this example, the observed mean difference is 25 ms, so the confidence interval around this mean difference score would be 25±(13.5 x $\sqrt{2}$) = 25±19.1.

Why is the denominator of Equation 9 nxJ (= 4x10 = 40 in this example) rather than the usual n (= 10 in this example) as would be the case if this were a pure between-subjects design? The reason for this further illustrates the different conclusions that can be made from a within-subjects design compared to a between-subjects design. In a purely between-subjects design, the confidence interval applies to a single condition mean. However, in this kind of mixed design, the confidence interval for the between-subjects factor—age—applies to the entire age curves rather than just a single mean. For this reason, the confidence interval is actually around an entire curve mean which is based on nxJ, or in this case, 40 observations. Again, this issue is most easily conceptualized by imagining the situation in which the within-subjects confidence interval is zero, and that the only uncertainty in the experiment is of age. The age uncertainty applies to an entire curve, not an individual mean; that is, once a given mean within a particular curve is known, the remaining three means in the curve are similarly known.

Confidence intervals around interaction effects

Often the nature of an interaction is a key factor underlying the conclusions that are made from some data set. Interactions with more than a single degree of freedom are the topic of a later section on contrasts. In this section, I will briefly describe how a one-degree-of-freedom interaction may be assessed as a single value plus a confidence interval rather than within the usual hypothesis-testing context.

Table 5 shows a hypothetical example of 2x2 design. The magnitude of the interaction may be computed as:

$$I = (M_{21} - M_{22}) - (M_{11} - M_{12})$$

which in this case is, I = 2.0. Suppose that the confidence interval around the individual mean is

Table 5. Hypothetical data from a 2 x 2 factorial design.

		Factor 1	
		Level 1	Level 2
Factor 2	Level 1	$M_{11} = 5$	$M_{21} = 8$
	Level 2	$M_{12} = 7$	$M_{22} = 12$

computed to be X (e.g., suppose X = 0.4 in this example). Thus, by Equation 7, the confidence interval around this interaction magnitude is,

$$I \pm x\sqrt{1^2 + 1^2 + 1^2 + 1^2} = I \pm 2X$$

which, in this example, would be 2.0 ± 0.8.

Asymmetrical confidence intervals

Thus far in the chapter I have been describing confidence intervals that are symmetrical around the obtained sample statistics (generally the sample mean). However, some circumstances demand asymmetrical confidence intervals. In this section, I will describe how to compute asymmetrical confidence intervals around three common statistics: variances, Pearson r's, and binomial proportions. In general, asymmetry reflects the bounded nature of the variable: variances are bounded at zero; Pearson r's are bounded at ±1).

Confidence intervals around variances

As described by Hays (1973, pp. 441-445) the confidence interval for a sample variance based on n observations (X_i's) with mean M, is:

$$CI = \begin{matrix}\text{(Upper limit):} & \dfrac{(n\text{-}1)\text{est}\sigma^2}{\chi^2(n-1;\ p(\text{upper limit}))} \\ \text{(Lower limit):} & \dfrac{(n\text{-}1)\text{est}\sigma^2}{\chi^2(n-1;\ p(\text{lower limit}))}\end{matrix}$$

Here, $\text{est}\sigma^2$ (or s^2 in Hays' notation) is the best estimate of the population variance computed by,

$$\text{est}\sigma^2 = \frac{\sum_{i=1}^{n}(X_{ij} - M)^2}{n-1} = \frac{\sum_{i=1}^{n}X_i^2 - nM^2}{n-1}$$

and p(upper limit) and p(lower limit) are the probability boundaries for the upper and lower limits of the confidence interval (e.g., 0.975 and 0.025 for a 95% confidence interval).

Suppose, to illustrate, that a sample of n = 100 scores produced a sample variance, $\text{est}\sigma^2 = 20$. The upper limit of a 95% confidence interval would be,

$$\frac{(100\text{-}1)(20)}{\chi^2(9,0.975)} = \frac{99x20}{73.36} = 26.99$$

while the lower limit would be,

$$\frac{(100\text{-}1)(20)}{\chi^2(9,0.025)} = \frac{99x20}{128.42} = 15.42$$

Confidence intervals around Pearson r's

The confidence interval around a Pearson r is based on Fisher's r-to-z transformation. In particular, suppose a sample of n X-Y pairs produces some value of Pearson r. Given the transformation,

$$z = 0.5\ln\left(\frac{1+r}{1-r}\right) \qquad \text{(Eq. 10)}$$

z is approximately normally distributed, with an expectation equal to

$$0.5\ln\left(\frac{1+\rho}{1-\rho}\right)$$

where ρ is the population correlation of which r is an estimate, and a standard deviation of

$$\sigma = \sqrt{1/(n-3)}$$

Therefore, having computed an obtained z from the obtained r via Equation 10, a confidence interval can easily be constructed in z-space as

$$z \pm \text{criterion } z$$

where the criterion z corresponds to the desired confidence level (e.g., 1.96 in the case of a 95% confidence interval). The upper and lower z limits of this confidence interval can then be transformed back to upper and lower r limits.

Suppose, for instance, that a sample of n = 25 X-Y pairs produces a Pearson r of 0.90, and a 95% confidence interval is desired. The obtained z is thus,

$$z = 0.5 \text{ x } \ln[(1+.90)/(1-.90)] = 1.472$$

which is distributed with a standard deviation of

$$\sqrt{1/(25-3)} = 0.213.$$

The upper and lower confidence interval limits in z-space are therefore

$$1.472+(.213)(1.96) = 1.890$$

and

$$1.472-(.213)(1.96) = 1.054.$$

To translate from z-space back to r-space, it is necessary to invert Equation 10, It is easily shown that such inversion produces,

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \qquad \text{(Eq. 11)}$$

The upper and lower confidence-interval limits may then be computed from Equation 11:

$$\text{upper limit: } r = \frac{e^{2x1.890} - 1}{e^{2x1.890} + 1} = 0.955$$

and

$$\text{lower limit: } r = \frac{e^{2x1.054} - 1}{e^{2x1.054} + 1} = 0.783$$

Thus, the 95% confidence interval around the original obtained r of 0.90 ranges from 0.783 to 0.955.

Confidence intervals around binomial proportions

To compute confidence intervals around binomial proportions, note first that the equation for the standard deviation of a proportion is,

$$\sigma = \sqrt{\frac{pq}{n}}$$

where p is the proportion, q is (1-p) and n is the number of observations.

Suppose now that we wish to compute the upper limit of a X% confidence interval. Call the corresponding criterion z, z_X (e.g., z_X = 1.64 for a 90% confidence interval, z_X = 1.96 for a 95% confidence interval, and so on). It follows then that the upper limit, U, for an X% confidence interval around some obtained proportion, p, can be written as,

$$U = p + \frac{1}{2n} + z_x\sigma = p + \frac{1}{2n} + z_x\sqrt{\frac{U(1-U)}{n}} \quad \text{Eq. 12}$$

where the factor (1/2n) is to correct for continuity, as the normal approximation to the binomial is most easily used in these computations. The equation for the lower limit, L, is the same except that the second plus sign in Equation 12 is replaced with a minus sign, i.e.,

$$L = p + \frac{1}{2n} - z_x\sigma = p + \frac{1}{2n} - z_x\sqrt{\frac{L(1-L)}{n}}$$

This equations for both U or L, can, after suitable algebraic manipulation, be written as standard quadratics of the form,

$$aU^2 + bU + c = 0$$

and,

$$aL^2 + bL + c = 0$$

where for both U and L, the values of a, b, and c can be computed as,

$$a = 1 + \frac{z_x^2}{n} \qquad \text{Eq. 13}$$

and,

$$b = -2p - \frac{z_x^2}{n} - \frac{1}{n} \qquad \text{Eq. 14}$$

and

$$c = p^2 + \frac{p}{n} + \frac{1}{4n^2} \qquad \text{Eq. 15}$$

The seemingly odd fact that the values of a, b, and c are the same for *both* U and L comes about because when, as part of the aforementioned algebraic manipulation, one squares the far-right term in Equation 12: the minus sign in the equation for L disappears and hence the equations for U and L become identical. Nevertheless, distinct values for both U and L emerge from the quadratic solution below.

A quadratic equation of the form,

$$aX^2 + bX + c = 0$$

has two solutions, which are computed as follows.

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \qquad \text{Eq. 16}$$

When the values of a, b, and c obtained by Equations 13, 14, and 15 are plugged into Equation 16, the two resulting solutions correspond to the U and L, the upper and lower limits of the confidence interval.

As an example, supposed that an obtained proportion of p = .96 is obtained based on n = 5 observations, and suppose one wishes to compute a 99% confidence interval around this obtained value of p = .96. The criterion z for a 99% confidence interval is z_X = 2.576. There is now sufficient information to compute the values of the quadratic-equation coefficients, a, b, and c via Equations 13-15. They are, a = 2.327, b = -3.447. and c = 1.124. Plugging these three values, in turn, into the Equation 16 leads to solutions—upper and lower limits—of U = 0.997 and and L = 0.484.

Homogeneity of Variance

Let us return to the standard, one-way, between-subjects ANOVA design, as exemplified the RT-as-a-function-of-caffeine example (see Figure 5). There is only a single MS (Error) in this design, in this case MS (Within) = 323. Computation of this single MS (Within) rests on the *homogeneity of variance assumption* which is this: Although the treatment in some experiment (caffeine variation in this example) may affect the population means, it does not affect population variances.

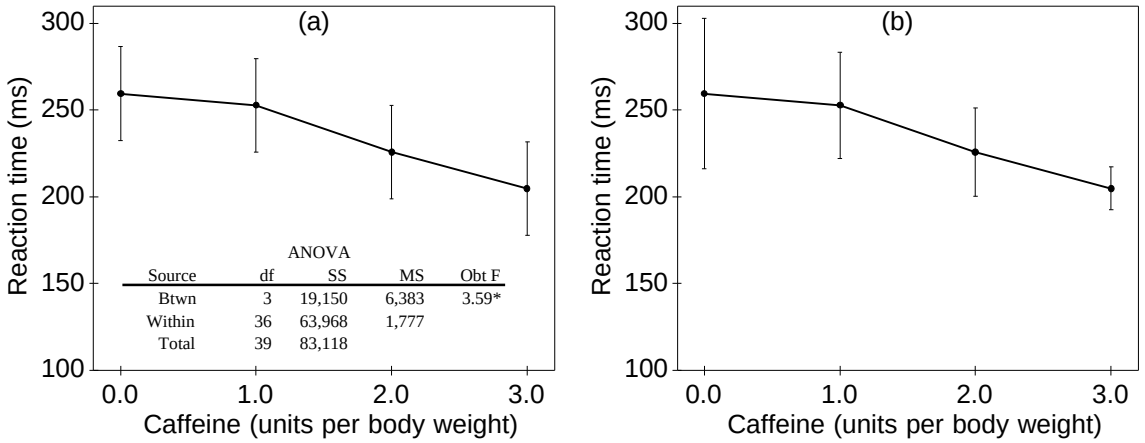

Figure 8. Caffeine data from a hypothetical between-subjects design similar to that of Figure 6. Homogeneity of variance is assumed (as usual) in the left panel, wherein an ANOVA is possible, and equal-sized 95% confidence intervals are shown. Homogeneity of variance is not assumed in the right panel. An ANOVA cannot be carried out; however the different-sized 95% confidence intervals represent the differently estimated variances in the different conditions.

Accordingly, there is assumed to be a single population variance, σ^2, that characterizes the populations corresponding to all levels of the independent variable. Although not apparent in the usual formulas, the MS (Within) is the weighted average of separate estimates of σ^2 obtained from each level of the independent variable[11].

The homogeneity of variance assumption, although almost invariably false, is necessary for carrying out an ANOVA. The consequences of violating the homogeneity of variance assumption to a mild degrees are not severe (see, e.g., Hays, 1973, pp. 481-483). The homogeneity of variance assumption is not necessary at all, however, for computing confidence intervals. In the following sections, I will touch on computation of confidence intervals in the absence of the homogeneity of variance assumption in several representative designs and, in the process, demonstrate the value of confidence intervals in illuminating the effects of the independent variable on condition variance as well as on condition mean.

Single-factor between-subjects designs

In a single-factor between-subjects design such as the one illustrated in Figure 5, the relevant LM equation is,

$$Y_{ij} = \mu + \alpha_j + e_{ij} \qquad \text{(Eq. 17)}$$

where Y_{ij} is the score for Subject i in Condition j, μ is the grand population mean, α_j is the effect of Treatment (Condition) j, and e_{ij} is an error associated with subject i in Condition j. Homogeneity of variance is reflected by the assumption that the e_{ij}'s are distributed normally with a mean of zero and a variance, σ^2, that is independent of j.

If the investigator is willing to forego an ANOVA, the homogeneity of variance assumption may be dropped in favor of the more general and realistic assumption that the independent variable affects condition variance as well as condition mean, i.e., that the variance of the e_{ij}'s in Equation 17 is σ^2_j for Condition j. To illustrate, let us return to the single-factor caffeine experiment whose results are depicted in Figure 5. Suppose, that the data from this experiment had turned out as depicted in Figure 8a. Making the standard homogeneity of variance assumption, a single confidence interval can be computed based on MS (Within) and displayed as shown.

Suppose that the homogeneity of variance assumption necessary for the ANOVA were dropped, and separate confidence interval's were computed for each condition by,

$$CI_j = \sqrt{\frac{\text{est}\sigma_j^2}{n_j}} \text{ crit } t(n_j - 1)$$

where j indexes condition. Here, $\text{est}\sigma_j^2$ is the estimate of Condition j's population variance, computed by

$$\text{est}\sigma_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - M_j)^2}{n_j - 1} = \frac{\sum_{i=1}^{n_j} x_{ij}^2 - T_j^2 / n_j}{n_j - 1}$$

[11] The weighting is by degrees of freedom. In the example at hand, there are equal n's and hence equal degrees of freedom in each condition.

where T_j, M_j and n_j are, respectively, the total of, mean of, and number of subjects in the jth condition (n_j = 10 for all conditions in this example). Note that whereas when assuming homogeneity of variance as in Figure 8a, the criterion t for the confidence interval is based on degrees of freedom within (36 in this example). When not assuming homogeneity of variance, the criterion t for the Condition-j confidence interval is based on (n_j-1) degrees of freedom, the number of degrees of freedom in Condition j.

These new confidence intervals—computed not assuming homogeneity of variance—are plotted in Figure 8b, which provides important intuitive pictorial information about the effect of caffeine on variance that is not available in the ANOVA of Figure 8a. In particular, it suggests that caffeine's effect on the variance should be construed as at least as important as caffeine's effect on the mean.

Multi-factor between-subjects designs

Considerations involving homogeneity of variance become more complex when more than a single factor is included in the design, as there are many configurations of variance homogeneity that could be assumed. For a two-factor, JxK design, the most coherent possibilities are as follows. (For simplicity, I assume equal n's in all conditions).

1. Complete homogeneity of variance is assumed. In this case, a single confidence interval can be computed, appropriate for each of the JxK conditions, based on (JxK)x(n-1) degrees of freedom within.

2. No homogeneity of variance is assumed at all. In this case, a confidence interval can be computed independently for the each of the JxK conditions. The confidence interval for the JKth condition is based on (n-1) degrees of freedom.

3. Homogeneity of variance can be assumed across the J levels of Factor 1 but not across the K levels of Factor 2. In this case, K confidence intervals are computed, one for each level of Factor 2, each based on Jx(n-1) degrees of freedom. The confidence interval for Level k of Factor 2 is appropriate for all J Factor-1 levels within Level k of Factor 2.

4. Conversely, homogeneity of variance can be assumed across the K levels of Factor 2 but not across the J levels of Factor 1. In this case, J confidence intervals are computed, one for each level of Factor 1, each based on Kx(n-1) degrees of freedom. The confidence interval for Level j of Factor 1 is appropriate for all K Factor-2 levels within Level j of Factor 1.

Single-factor within-subjects designs

In a single-factor within-subjects design illustrated in Figure 6, the issue of homogeneity of variance is somewhat complicated. The relevant LM equation is,

$$Y_{ij} = \mu + \beta + \alpha_j + \gamma_{ij}$$

where Y_{ij}, and α_j are as in Equation 17, β_i is an effect of Subject i, and γ_{ij} is an "interaction term" unique to the Subject i x Condition j combination. Homogeneity of variance in this design is the assumption that the γ_{ij} terms are all distributed normally with a variance of σ^2. Dropping the homogeneity of variance assumption would allow the variance of the γ_{ij} terms to have different variances σ^2_j for the different conditions, j.

Estimation of the separate σ^2_j's is described by Loftus and Masson (1994), p. 484 and in their Appendix B. Unlike the corresponding between-subjects situation described in the previous section, such separate estimation is sufficiently involved that I will not re-describe it here. Moreover, the procedure entails potential estimation problems described by Loftus and Masson (that are exacerbated by small sample sizes). For this reason, it is not recommended that this procedure be used unless there is very good reason to do so.

Multi-factor within-subjects designs

Many of the same considerations that apply to multi-factor between-subjects designs apply similarly to multi-factor within-subjects designs. Consider for example a J (Factor 1) x K (Factor 2) x n (subjects) design. Although, as just noted, it is somewhat tedious to estimate different variances, σ^2_j, of the γ's corresponding to the J different levels within a given factor, it is simple to estimate values of γ if they are presumed different for different levels of Factor 2, but, within each level of Factor 2, the same for all levels of Factor 1: One need only apply Equation 9 separately and independently for each level of Factor 2. (And, of course, the same logic applies reversing Factors 1 and 2).

To illustrate, suppose that again the effect of caffeine on RT is under consideration. In this hypothetical example, Factor A is amount of caffeine (which again can be one of four levels), while Factor B is amount of sleep deprivation which is either 1 or 24 hours. Suppose n = 10 subjects participate in each of the 8 caffeine x deprivation conditions. Assume for the sake of argument that the three error terms—the interactions of subject x caffeine, subject x deprivation, and subject x caffeine x deprivation—are approximately the same. Using the logic described above, the investigator could compute a single confidence interval using

the combined error term, which would be based on 9 (degrees of freedom for subjects) x 7 (degrees of freedom for the 8 conditions) = 63 degrees of freedom (or alternatively, 9x3 + 9x1 + 9x3x1 = 63 degrees of freedom if one prefers to think in terms of adding the degrees of freedom from the three separate error terms).

Suppose alternatively, that the investigator were suspicious that the effect of caffeine was less consistent over subjects with 24 hours of sleep deprivation than with 1 hour of sleep deprivation. Again, foregoing a standard ANOVA, the investigator could essentially view the design as comprising two separate experiments—one involving the effect of caffeine on RT following one hour of sleep deprivation and the other involving the effect of caffeine on RT following 24 hours of sleep deprivation. Two confidence intervals could then be computed, each based one of these two "separate experiments"—i.e., each based on the subject x caffeine interaction within one of the sleep-deprivation levels—and each based on 9x3 = 27 degrees of freedom.

Planned Comparisons (Contrasts)

Psychological research, along with the analysis of psychological data, varies widely in the degree of quantitative sophistication that is used. At or near one end of this continuum is the use of mathematical process models to generate quantitative predictions for the summary statistics obtained in an experiment (and in some cases, distributions of the raw data as well) (see, e.g., Myung & Pitt's chapter, this volume). At, or near the other end of the continuum is NHST used to evaluate verbally presented hypotheses wherein the only mathematical model is some form of the standard linear model. The use of planned comparisons falls somewhere in the middle. Planned comparisons provide an organized and systematic means of accounting for variability between conditions in some experiment.

The formal logic and details of the use of planned comparisons is presented Hays (1973, pp. 584-593). The basic logic of a planned comparison is as follows

1. Some hypothesis about what the pattern of population means looks like is used to generate a set of numbers called weights—one weight corresponding to each condition in the experiment. The general idea is that the pattern of weights over conditions correspond to the pattern of population means that is predicted by the hypothesis. (It is important to realize that, unlike a mathematical model designed to generate an exact quantitative prediction for each condition, each individual weight of a planned comparison need not bear any particular relation to its corresponding sample mean. Rather, it is the *pattern* of weights that should correspond to the predicted pattern of means. In most applications of planned comparisons, the weights must sum to zero, in which case the comparison is conventionally referred to as a *contrast*.
2. The correlation (Pearson r^2) between the hypothesis weights and the sample means is computed. This Pearson r^2, like any Pearson r^2, is interpreted as the percent of variance between conditions, i.e., the percent of SS (Between), that is accounted for by the hypothesis.
3. Accordingly the product of the Pearson r^2 and SS (Between) is interpretable as a sum of squares. This sum of squares is based on one degree of freedom.
4. Within the context of NHST two null hypotheses can be tested. The first, which I label a "uselessness null hypothesis" is that the correlation between the hypothesis weights and the condition population means is 0.0 (informally, that the hypothesis is useless as a descriptor of reality). The second, which I label a "sufficiency null hypothesis" is that the correlation between the hypothesis weights and the condition population means is 1.0 (informally, that the hypothesis is sufficient as a descriptor of reality).

An Example of the Use of Planned Comparisons

Suppose that an investigator is studying factors that influence attitude change. The general paradigm is this. Subjects listen to a speaker who describes the benefit of a somewhat controversial issue, specifically clearcutting in national forests. Following the speech, the subjects rate the degree to which they favor the speaker's position on a scale from 1 ("don't agree at all") to 7 ("agree fully"). In an initial experiment, the effect of speaker affiliation is investigated. In J = 5 conditions, subjects are provided either (1) no information, or information that the speaker is a member of (2) the Sierra Club, (3) the Audubon Society, (4) the timber industry, or (5) the paper industry. The Conditions are summarized in Table 6, Panel A.

Suppose the investigator wishes to test a hypothesis which is the conjunction of the following two assumptions. First, knowing something about the speaker leads to more attitude change than knowing nothing at all. Second, attitude change is greater for speakers whose affiliated organization

Table 6. Data from a hypothetical experiment in which attitude change (rating) is measured as a function of the perceived affiliation of the speaker. Panel A provides original data plus two successively constructed sets of weights: The W(2)'s are deviation scores obtained from the W(1)'s. Panel B shows the ANOVA results for the contrast and for the residual.

A. Means (M_j's) and Construction of Weights (W_j's)

Speaker information	Mj	Wj(1)	Wj(2)
None	2.25	0	-1.20
Sierra Club	6.05	2	0.80
Audubon Society	5.50	2	0.80
Timber industry	3.70	1	-0.20
Paper industry	2.90	1	-0.20

B. ANOVA

Source	df	SS	MS	F	%var = r^2
Between	4	215.7			
Hypothesis	1	194.6	194.6	19.86	0.902
Residual	3	21.1	7.0	0.72	0.098
Within	95	931.0	9.8		

is perceived to oppose the expressed opinion (i.e., the Sierra Club and the Audubon Society are perceived to oppose clearcutting) than for speakers whose affiliated organization is perceived to support the expressed opinion (i.e., the timber and paper industries are perceived to support clearcutting).

To assess the viability of this hypothesis, the sample means are plotted in Figure 9 along with the confidence intervals. The pattern of observed sample means appears to roughly bear out the hypothesis: The "None" condition produces the lowest mean persuasion value, the Sierra-Club and Audubon-Society values are highest, and the timber and paper industry conditions are intermediate.

To acquire a quantitative handle on this apparent confirmation of the hypothesis, a planned comparison is carried out. To do this, the investigator's first job is to create a set of weights that reflects the hypothesis described above. The first step in doing so is to create weights ignoring for the moment the constraint that the weights must sum to zero. The simplest such weights would assign zero to the "None" condition, 2's to the Sierra Club and Audubon Society conditions, and 1's to the timber industry and paper industry conditions. These weights are provided in the Table-6, Panel-A column labeled "$W_j(1)$". The next step is to preserve the pattern produced by this set of weights but make the weights add to zero. This is easily accomplished by computing the mean of the $W_j(1)$'s, which is 1.2, and subtracting that mean from the $W_j(1)$'s to generate a set of *deviation scores* which, while preserving the pattern of the $W_j(1)$'s are, of course, guaranteed to add to zero. The resulting final weights are provided in the fourth column, labeled "$W_j(2)$". (It is worth pointing out that this any-numbers-then-make-deviation-scores trick is quite useful for generating weights in *any* situation.)

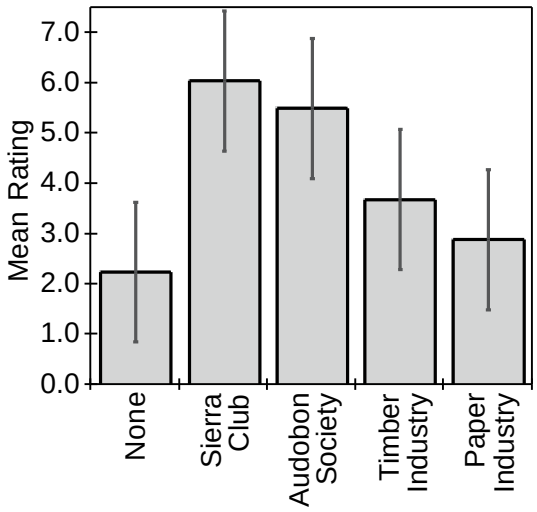

Figure 9. Data from a hypothetical experiment in which attitude change (rating) is measured as a function of the perceived affiliation of the speaker. The error bars are 95% confidence intervals.

Percent of between-condition variance accounted for by the hypothesis

As noted, a basic goal is to compute the Pearson r^2 between the sample means and the weights corresponding to the hypothesis. Although this could easily be done using the standard Pearson-r^2 equation, it is more instructive, within the context of planned comparisons, to do the computation via a somewhat different route. In particular, a sum of squares due to the hypothesis may be computed using the equation,

$$SS(\text{Hypothesis}) = \frac{n\left(\sum_{j=1}^{J} M_j W_j\right)^2}{\sum_{j=1}^{J} W_j^2} \qquad \text{(Eq. 18)}$$

where n is the number of subjects in each condition (n = 20 in this example). Applying Equation 18 to the present data produces SS (Hypothesis) = 194.6, shown in the Table-6, Panel-B

ANOVA table. The ratio of SS (Hypothesis) to SS (Between) is 194.6/215.7 = 0.902 which is the sought-after Pearson r^2 between the W_j's and the M_j's.

This sum of squares representing, as it does, a single pattern of variation across the five conditions, is based on one degree of freedom. By subtraction one can compute the portion of SS (Between) that is not accounted for by the hypothesis: This is 215.7-194.6 = 21.1, a value which is referred to as SS (Residual). SS (Residual) represents all forms of variability other than that engendered in the original hypothesis and is based on 3 degrees of freedom = 4 (df (Between)) - 1(df (Hypothesis)).

Mean squares can be computed in the normal fashion based on sums of squares and degrees of freedoms due to the hypothesis and the residual; these mean squares are in the column labeled "MS." If one is inclined to work within the NHST framework, then these mean squares are used to test two null hypotheses[12].

A "uselessness" null hypothesis

The Pearson r^2 of 0.902 shown in Table-6 "%var = r^2" column is the r^2 between the sample means (M_j's) and the weights (W_j's). As in any situation involving unknown population parameters, it would be of more interest to address the Pearson r^2 between the W_j's and the population means, i.e., the μ_j's. Two null hypotheses are relevant to this issue. The first is the null hypothesis that the Pearson r^2 between the W_j's and the μ_j's is zero, i.e., that the hypothesis is useless as an account of SS (Between). If this were true, then the MS (Hypothesis) as shown in Table 6 is an estimate of MS (Within) and a standard F test can be carried out wherein F (dfH, dfW) = MS (Hypothesis)/MS (Within). As indicated in Table 6, this F, which is 19.86 is statistically significant, thereby allowing rejection of this "uselessness null hypothesis".

A "sufficiency" null hypothesis

The second null hypothesis is that the Pearson r^2 between the W_j's and the μ_j's is 1.0, i.e., that the hypothesis is sufficient to account for SS (Between). Testing this null hypothesis entails an F ratio of MS (Residual) against MS (Within). In Table 6, it can be seen that the resulting F(3, 95) is 0.72 which is, of course, nonsignificant.

[12] Some terminology glitches arise here. I want to emphasize that the term "hypothesis" when used alone refers to a form of an alternative hypothesis. The term "null hypothesis" refers to the two specific quantitative hypotheses that will be described.

Reminder of problems with NHST

It is of course necessary to bear in mind that these uses of NHST carry with them all of the problems with NHST described earlier. In particular, an outcome such as the one portrayed in Table 6—that the hypothesis is significant, but the residual is not—should be accompanied by a number of caveats, the most important of which is that failure to reject the "sufficiency null hypothesis" does not mean that the sufficiency null hypothesis is correct. Indeed in the present example it should set off alarm bells that only 90% of the between-condition variance is accounted for by the hypothesis. These are the same alarm bells that should be set off by the relatively large confidence intervals that are depicted in Figure 9.

Planned Comparisons of Linearity

A frequent use of planned comparisons is to test a hypothesis of *linearity*. Suppose, to illustrate, that an investigator is studying the effect of audience size on the degree of stage fright suffered by a public speaker (e.g., Jackson & Latané, 1981). In a hypothetical experiment, subjects give prepared speeches to audiences whose sizes are, in different conditions, 3, 6, 12, 20, or 29 people. Following the speech, a subject indicates degree of stage fright that he or she has experienced on a scale ranging from 0 ("not frightened at all") to 7 ("terrified"). A between-subjects design is used with n = 15 subjects participating in each of the J = 5 audience-size conditions. The data from this experiment, shown in Figure 10, are provided in Table 7, which is organized like Table 6.

Suppose the investigator wishes to test the hypothesis that stage fright, as measured by the rating, increases linearly with audience size; thus, the best linear fit is provided in Figure 10 along with the data points. It appears that a linearity hypothe-

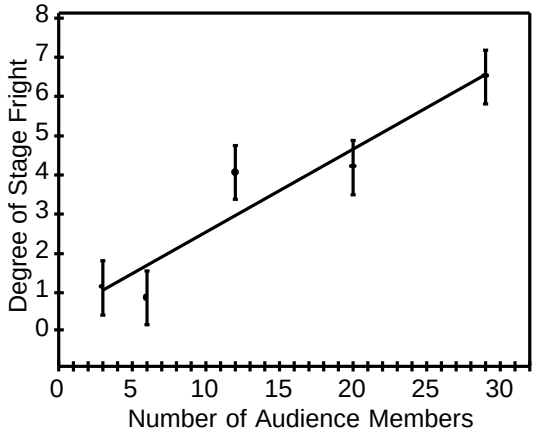

Figure 10. Data from a hypothetical experiment in which stage fright (rating) is measured as a function of audience size. The error bars are standard errors.

Table 7. Data from a hypothetical experiment in which stage fright (rating) is measured as a function of audience size. Panel A provides original data plus three successively constructed sets of weights: The W(2)'s are deviation scores obtained from the W(1)'s, and the W(3)'s are scaled W(2)'s (scaling designed to render the contrast in "natural units" as described in the text). Panel B shows the ANOVA results for the contrast and for the residual.

A. Means (M_j's) and construction of Weights (W_j's)

Audience size	M_j	$W_j(1)$	$W_j(2)$	Wj(3)
3	1.100	3	-11	-0.0244
6	0.833	6	-8	-0.0178
12	4.033	12	-2	-0.0044
20	4.167	20	6	0.0133
29	6.500	29	15	0.0333

B. ANOVA

Source	df	SS	MS	F	%var = r^2
Between	4	336.7			
Hypothesis	1	305.1	305.1	64.0**	0.906
Residual	3	31.6	10.6	2.11 ns	0.094
Within	70	514.5	7.3		

sis is roughly confirmed.

The first task in carrying out the linearity planned comparison is to generate weights that are linearly related to audience size. This enterprise is complicated slightly because the audience-size levels (3, 6, 12, 20, 29) are not evenly spaced. The simplest way of coping with this complication is to use the trick described earlier, and begin by selecting appropriate weights—in this case, weights that are linear with audience size—without concern about whether they add to zero. A simple and suitable candidate for such weights are the audience sizes themselves, as indicated in the Table-7 column labeled "$W_j(1)$". As in the previous example, this pattern of weights can be made to add to zero by transforming the $W_j(1)$'s to deviation scores. The resulting weights are provided in the column, labeled "$W_j(2)$". (The "$W_j(3)$" column will be described in the next section). The remainder of the process is exactly as was described in the previous example: The $W_j(2)$'s are plugged into Equation 18 to find the SS (Hypothesis), SS (Residual) is found by subtraction, the percents of the SS (Between) accounted for by the Hypothesis and Residual are computed and the uselessness and sufficiency null hypothesis tests are carried out. These results are shown in the Table-7, Panel-B ANOVA table.

A Contrast as a Dependent Variable: Scaling the W_j's

Let us examine Equation 18 more closely. The heart of the equation is in the term that constitutes the actual contrast,

$$\text{Contrast} = C = \sum_{j=1}^{J} M_j W_j \qquad \text{(Eq. 19)}$$

The larger is C, the "better" the hypothesis may be assumed to be. Often it is useful to view C as a dependent variable in the experiment. This strategy is particularly advantageous if the contrast has easily interpretable or "natural" units. A very simple example of such use occurs when the weights are all zero except for a "1" and a "-1" in which case the contrast is interpretable as a *difference score*.

However, more sophisticated uses of contrasts as a "natural" dependent variable can be engineered. Before providing an example of how this might be done, it is critical to point out the *scalability property* of the W_j's. To understand this property, note that the denominator of Equation 18 serves to eliminate any effect of *scaling* the weights. Suppose that an investigator has chosen some suitable set of weights, $W = (W_1, W_2,..,W_J)$ and a SS (Hypothesis) were computed via Equation 18. Now suppose that an alternative, set, $W' = kW = (kW_1, kW_2,\ldots,kW_J)$, were used where k is some nonzero constant. Applying Equation 18 to W' would yield a factor of k^2 both numerator and denominator compared to using the original W. Therefore, the k^2's would cancel and the same SS (Hypothesis) and r^2 would emerge. In short, once one has chosen a suitable set of weights, any other scaled set is equally suitable.

An investigator can use this fact to his or her advantage to scale weights in such a way that the contrast is expressed in some form of natural units. An obvious example of this sort of procedure is when a linear hypothesis is under investigation, as in the stage-fright example depicted in Figure 10 and Table 7. In particular, a "natural unit" for the contrast would be the *slope* of the function relating stage-fright rating to audience size. How might the Table-7 weights be scaled to accomplish this?

The $W_j(2)$ weights from Table 7 are already scaled in units of audience size; they are just shifted so as to constitute deviation scores. Thus, the slope of the audience-size function may be computed using the standard regression equation,

$$\text{slope} = \frac{5\sum_{j=1}^{5} M_j W_j - \left(\sum_{j=1}^{5} M_j\right)\left(\sum_{j=1}^{5} W_j\right)}{\sum_{j=1}^{5} W_j^2 - \left(\sum_{j=1}^{5} W_j\right)^2}$$

or, because the W_j's must sum to zero,

$$\text{slope} = \frac{\sum_{j=1}^{5} M_j W_j}{\sum_{j=1}^{5} W_j^2}$$

This in turn means that if the original weights (i.e., the $W_j(2)$ weights from Table 7) are scaled by a factor of $1/\Sigma W_j^2 = 1/450$, then a set of weights will emerge that will produce as a contrast the slope of the function. It is these scaled weights that are labeled $W_j(3)$ in Table 7. Applying Equation 19 to the M_j's and the $W_j(3)$'s from Table 7, yields C = 0.213 which is the slope of the audience-size function.

Confidence intervals around Contrasts

One can also compute a confidence interval around the observed value of C. Such computation is straightforward. As is well known, and indicated in Equation 7, any linear combination of means, as in Equation 19 has a variance of

$$\sigma_C^2 = \sigma_M^2 (W_1^2 + W_2^2 + ... + W_j^2)$$

where σ_M^2 is the standard error of the mean (it is necessary, of course, to assume homogeneity of variance here). Because σ_M^2 is estimated by [MS (Within)]/n, the standard error of C may be computed as,

$$SE = \pm\sqrt{\frac{MSW}{n}(W_1^2 + W_2^2 + ... + W_j^2)} \quad \text{(Eq. 20)}$$

and any desired-size confidence interval may be computed by multiplying Equation 20 by the appropriate criterion t(dfW).

Recall that the contrast from the Table-7 $W_j(3)$'s was C = slope = 0.213. Applying Equation 18 to the Table-7 MS (Within) and the $W_j(3)$'s yields a 95% confidence interval of 0.066. In short, one may summarize the stage-fright data by stating that the slope of the audience-size function is 0.213 with a 95% confidence interval of 0.066.

Using Planned Comparisons in Within-Subjects Designs

Planned comparisons can be used in within-subjects designs much in the same way that they can be used in between-subjects designs.

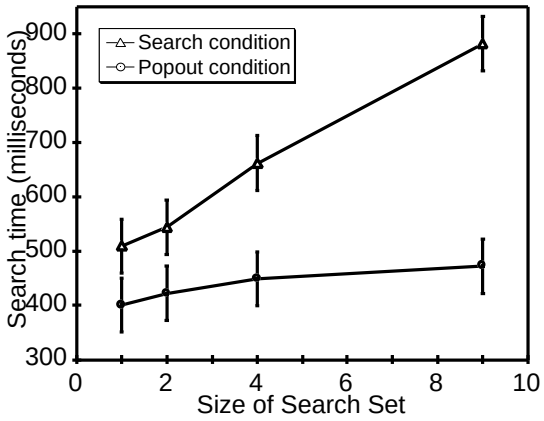

Figure 11. Data from a hypothetical experiment in which search time (RT) is measured as functions of set size and whether search is required. The error bars are 95% confidence intervals.

Example: Visual search and "popout"

As an example, consider a visual search task in which the subject's task is to determine whether or not some target stimulus is present amongst some set of distractors. Suppose that two conditions are constructed: a "search" condition in which it is predicted that the subject will have to search serially to make the decision and a "popout" condition in which it is predicted that the subject will be able to process all members of the stimulus array in parallel. Search set size is also varied, consisting of 1, 2, 4, or 9 items. The design is entirely within-subjects, and the 8 conditions defined by 2 (search/popout) x 4 (set size) are presented randomly to each of n = 9 subjects over a long series of trials.

The date from this hypothetical experiment (means plus confidence intervals) are shown in Figure 11. It is clear that RT increases with set size in the search condition whereas RT is relatively low and flat in the popout condition. These means are reproduced numerically in Table 8, Panel A. Table 8, Panel B shows a standard ANOVA table for these data. F-ratios have been computed for the standard three factors—effects of set size, search/popout, and the interaction. As one would surmise from Figure 11, all three of these effects are highly significant.

Testing the hypothesis with planned comparisons

Of primary interest is testing the prediction of the hypothesis described above, i.e., that RT should increase linearly with set size in the search condition, but should be flat (and presumably low) in the popout condition. I now describe two planned comparisons that are suitable for doing

this. I should note that the overall ANOVA shown in Table 8, Panel B was provided for expositional purposes only: When the investigator carries out planned comparisons, the overall ANOVA is generally not necessary.

Table 8. Panel A: Original search-time data. Panel B: ANOVA results. Panels C and D: Contrasts described in the text.

A. Original Data (in milliseconds)

	Number of Items in the Search Set			
	1	2	4	9
Search	509	544	662	882
Popout	400	422	449	472

B. Overall ANOVA

Source	df	SS	MS	F
Subjects (S)	7	219,325		
Conditions (C)	7	1,432,963		
Set Size (Z)	3	471,767	157,256	38.68
Search/Popout (P)	1	729,957	729,957	136.18
Z x P	3	231,239	77,080	13.80
S x Z	21	85,386	4,066	
S x P	7	37,522	5,360	
S x Z x P	21	117,336	5,587	
S x C	49	240,244	4,903	
TOTAL	63	1,892,532		

C. Contrast from total SS (Between cells)

	1	2	4	9
$W_{jk}(1)$	1	1	1	1
$W_{jk}(2)$	-1.5	-0.5	1.5	6.5
	-1.5	-1.5	-1.5	-1.5

Contrast ANOVA

Source	df	SS	MS	F	%var = r^2
Conditions	7	1,432,963			
Hypothesis	1	459,456	459,456	93.71	32.1
Residual	6	973,506	162,251	33.09	68.9

D. Contrast from Interaction SS only

	5	6	8	13
$W_{jk}(1)$	11	10	8	3
$W_{jk}(2)$	-3	-2	0	5
	3	2	0	-5

Contrast ANOVA

Source	df	SS	MS	F	%var = r^2
Interaction	3	231,239			
Hypothesis	1	229,907	229,907	46.89	99.4
Residual	2	1,333	666	0.14	0.6

A planned comparison based on total between-cell variance

The first planned comparison is shown in Table 8, Panel C. As with the previous examples, I use a two-part process to generate the appropriate W_{jk}'s. The $W_{jk}(1)$'s are constructed without regard to making the W_{jk}'s add to zero, while the $W_{jk}(2)$'s are the $W_{jk}(1)$ deviation scores.

This procedure illustrates an important point: When carrying out this kind of a planned comparison using a two-factor design (whether it be within-subjects as in the present example, or between-subjects) the row x column design structure becomes relevant only as a mnemonic aid. From the perspective of the planned comparison, the experiment is simply viewed as containing JxK different conditions, and the W_{jk}'s must add to zero across all JxK conditions. There are, for the moment, no other constraints on the W_{jk}'s.

The statistical results of this procedure are shown at the bottom of Panel C labeled "Contrast ANOVA." The top source of variance is from between conditions (i.e., the Panel-B component labeled "Conditions", and is based on 7 degrees of freedom. The SS (Hypothesis), computed via Equation 18, is the next variance component, and finally, as in previous examples, SS (Residual) is computed by subtraction. Note from the rightmost column that the SS (Hypothesis) accounts for only 32% of the between-conditions variance. Both the hypothesis and the residual are highly significant.

A planned comparison based on interaction variance

The planned comparison described in the previous section had a certain degree of arbitrariness about it. Essentially, the main prediction under investigation was that there should be a particular type of interaction between set size and search/popout; the main effects of the two variables were of secondary importance. For example the large percent of between-condition variance not accounted for by the hypothesis comes about because, as is evident in Figure 11, the search condition RT is greater than the popout condition RT even in the set size = 1 conditions; the arbitrary choice in the Panel-C contrast was to assume these two conditions to be equal.

Accordingly, it would be useful to carry out a planned comparison that investigated the role of the particular expected interaction. The resulting contrast is constructed in Table 8, Panel D. The goal here is to maintain the hypothesized interac-

Table 9. Additional information for the visual-search data. Panels A and B show the original data along with $W_{jk}(1)$ and $W_{jk}(2)$ from Table 7. The $W_{jk}(3)$'s are scaled $W_{jk}(2)$'s (scaling designed to render the contrast in "natural units" as described in the text). Panel C: Values of the contrast for 8 subjects along with the mean and 95% confidence interval of the 8 contrast values.

A. Original Data (in milliseconds)

	Number of Items in the Search Set			
	1	2	4	9
Search	509	544	662	882
Popout	400	422	449	472

B. Contrast from Interaction SS only

$W_{jk}(1)$	5	6	8	13
	11	10	8	3
$W_{jk}(2)$	-3	-2	0	5
	3	2	0	-5
$W_{jk}(3)$	-0.070	-0.053	0.000	0.132
	0.079	0.053	0.000	-0.132

C. Contrast Values for Individual Subjects

Subject (k)	C_k
1	5.4
2	49.9
3	29.2
4	44.4
5	57.6
6	40.5
7	23.4
8	60.7
Mean	Confidence Interval
38.89	15.61

tion pattern in the eventual contrast, but to eliminate main effects. The resulting contrast shown in Panel D accomplishes this; note that each row and column of the final contrast (i.e., $W_{jk}(2)$) sums to zero, so only interaction variance is reflected. The interaction variance remains specifically that RT is positively linear with set size for search and negatively linear with set size for popout[13].

The ANOVA relevant to this contrast is shown at the bottom of Panel D. The top source of variance is from the interaction (i.e., the Panel-B component labeled "ZxP") and is based on 3 degrees of freedom. The interaction-only contrast accounts for over 99% of this interaction variance, and the small residual is not statistically significant.

Using a contrast as a dependent variable

It is instructive to once again illustrate how a contrast may be translated into "natural units" via suitable scaling of the weights. In the present example, a useful "natural unit" for the contrast would be the difference between the search and the popout slopes. To do this, we work from the weights in Table 8, panel D, where the interaction variance only is at issue. Again the $W_{jk}(2)$ weights are already scaled to set size. Using much the same logic entailed in scaling the weights in the stage-fright example, and noting the constraints on the $W_{jk}(2)$ weight pattern, it can be shown that the appropriate scaling factor is $2/\Sigma W_{jk}^2$ where the sum is over all 8 conditions. The resulting weights, $W_{jk}(3)$, are shown in Table 9, Panel B. (Note that Panel A, along with part of Panel B re-presents relevant information from Table 8). Table 9, Panel C shows the contrasts (C_k's) that result for each subject, k. Thus, the contrast value for each subject—which, recall, has been designed to be the difference between the two slopes for that subject—can be treated as a standard dependent variable. The mean and 95% confidence interval shown at the bottom of Table 9, Panel C are computed from the C_k's directly.

Multiple Planned Comparisons

Multiple planned comparisons may be carried out on the same data set by generating multiple sets of weights, presumably from multiple hypotheses, and iterating through steps 1-4 described at the beginning of this section. Any two contrasts (along with the hypotheses that generated them) are independent of one another if and only if the Pearson r^2 between the two sets of weights is equal to zero. In practice, because any set of weights sums to zero, the Pearson r^2 between the two sets of weights is equal to zero if and only if the sum of the cross-products of the two sets of weights is equal to zero.

Percent Total Variance Accounted For (ω^2)

In correlational studies the primary dependent variable is a Pearson r^2. Every psychologist realizes that a Pearson r^2 represents the percent of

[13] That the hypothesis includes "negatively linear for popout may elicit some confusion because the original hypothesis predicted no set-size effect for popout. It must be borne in mind, however, that this contrast applies to the interaction variance only, which implies no main effect for set size, and which in turn implies canceling set-size effects for search and popout.

Table 10. Data from a hypothetical experiment in which the effect of Vitamin-C dosage on cold durations is examined. Panel A: Original data (10,000 subjects). Panel B: ANOVA results.

A. Original Data (n = 10,000/Condition)

Amount of Vitamin C (gms)	Mean days with colds
2	9.79
3	9.72
4	9.56

B. ANOVA

Source	df	SS	MS	F	Crit F
Between	2	290	145	16.12	3.00
Within	29,997	89,991	9		
Total	29,999	90,281			

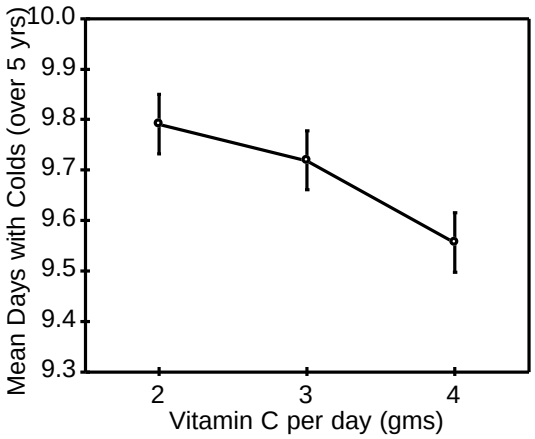

Figure 12. Data from a hypothetical experiment in which the effect of Vitamin-C dosage on cold durations is examined. Note the limited range of the ordinate. The error bars are 95% confidence intervals.

variance in some predicted variable, Y, accounted for by variation in some predictor variable, X.

Given the overwhelming prevalence of variance-accounted for measures such as Pearson r^2 in correlational research, it is puzzling that there is little use of the equivalent measures in experimental research. These measures, termed ω^2, are generally applicable to any ANOVA-type design and are, essentially, the percent of total variance in the experiment accounted for by variation in the independent variable. Computation of ω^2 is particularly useful in practical situations where the emphasis is on the effect's real-world significance (as opposed to its statistical significance). Hays (1973, pp. 417-424; 484-491; and 512-514) provides formal analyses of ω^2 for several experimental designs. I will briefly illustrate its use in a between-subjects, one-way ANOVA situation.

Vitamin C and colds

Suppose that an investigator is interested in determining whether variation in vitamin-C dosage affects the amount of time a person is afflicted with colds. In a hypothetical experiment, subjects in three different double-blind conditions, are provided, respectively, 2 gm, 3 gm, or 4 gm of vitamin C per day for five years and the number of days on which each subject considers him or herself to have a cold is recorded. A very large sample size is used: n = 10,000 subjects per condition. The data are provided in Table 10, Panel A and Figure 12, both of which make it clear that there is a highly significant, decreasing effect of Vitamin C on number of days with colds.

A closer inspection of the data, however, raises serious doubts about Vitamin C's efficacy: The absolute decrease in cold days is miniscule, falling from about 9.8 days to about 9.6 days as Vitamin C dosage is doubled from 2 to 4 grams. The reason that such a small effect is so highly significant is, of course, that the n = 10,000 subjects per condition confers an enormous amount of statistical power, and therefore even a tiny effect of the independent variable will be detected.

How much vitamin C should you take?

There are 30,000 data points in this experiment. As indicated in Table 10, Panel B, the total variance is SS (Total) = 90,281, of which only SS (Between) = 290 is from between conditions. Thus, essentially, only 290/90,281 or about 0.32% of the total variance is attributable to variation in vitamin C.

More precisely, because part of SSB is attributable to random error, the appropriate computation is somewhat more complex. The reader is referred to the Hays references provided above for the formal logic. The end result is that to estimate the percent of total variance attributable to variation in the independent variable, one uses the equation,

$$\text{est } \omega^2 = \frac{SS(\text{Between}) - (J-1)xMS(\text{Within})}{SS(\text{Between}) + SS(\text{Within}) + MS(\text{Within})}$$

which, in the present example is 0.30%. The inescapable implication is that vitamin C accounts for only a tiny percentage of total variability in cold days. The practical conclusion is that if one wishes to cut down on colds, there are probably many other more important variables to which one

should pay attention than the amount of vitamin C one takes.

This ω^2 computation places the large statistical significance found in this experiment in a somewhat harsher light, and serves to underscore the important difference between determining that an effect exists on the one hand (i.e., the high statistical significance implied by $F > 16$) and evaluating the effect's importance on the other hand (i.e., the minute practical significance implied by $\omega^2 < 1\%$).

A caveat

In the example I have just provided, the consumer of the research is presumably most interested in very practical considerations: Specifically, in deciding whether to go through the hassle and expense of taking large doses of Vitamin C it is useful to understand the magnitude of the expected reward in terms of cold relief. However, one might be interested in a separate question altogether, namely investigating the relation between vitamin C and colds strictly from the perspective of addressing some biological question. In such an instance, the relation between Vitamin C and cold reduction, no matter how small, could potentially be of intense interest.

Model Fitting

Myung & Pitt (this volume) describe the use of mathematical models in some detail. It is not my intent to reiterate what they have already said. Rather, in the spirit of the content of the present chapter, I will provide suggestions for understanding and presentation of some data/model combination.

Finding and Presenting Optimal Parameter Values

A simple and familiar example of fitting a mathematical model is in the context of linear regression, wherein some variable, Y is assumed to be linearly related to some other variable, X. Typically, some number of XY pairs constitute the data. Here two parameters must be estimated: the slope and the intercept of the assumed linear function relating Y to X. The standard equations for determining the best-fitting slope and intercept are based on the proposition that by "best" is meant the slope and intercept values that produce the smallest total squared error between the observed and predicted Y values.

Like a regression model, typical mathematical models have parameters. The main difference between fitting a simple regression model and a typical mathematical model is that the former has an analytical solution[14], while the latter usually do not. Even a model that is closely related to a linear model—an exponential growth to an asymptote model, expressed by the equation

$$Y = A(1 - e^{cX})$$

with two parameters c, the exponential decay rate and A, the asymptote—does not have an analytical solution. To find the best-fitting parameter values, some sort of search procedure is needed whereby candidate parameter sets are systematically evaluated, and the approximate best-fitting set is determined.

When carrying out such a search, it is necessary to decide what is meant by "best." Typically, one of three criteria are used to find the the parameter set that (1) minimizes total squared error between observed and predicted data points, or (2) minimizes the χ^2 between the observed and predicted frequencies, or (3) maximizes the probability of the data values given a particular parameter set (i.e., maximum likelihood techniques).

Fit quality expressed in intuitive units

My concern here is not with which technique is used—discussions of this issue may be found in many mathematical methods texts, e.g., Atkinson, Bower, & Crothers, 1965, Chapter 9—but with how the results of the search are presented. In particular, I recommend that, however the best-fitting parameter set is *found*, the quality of the fit be presented as root-mean-square-error (RMSE), which is obtained by,

$$\text{RMSE} = \sqrt{\frac{\sum_j (M_j - P_j)^2}{\text{degrees of freedom}}}$$

where the sum is over j experimental conditions, M_j and P_j are observed and predicted results in condition j, and degrees of freedom is degrees of freedom, which is approximately and most easily computed as the number of fitted data points minus the number of estimated parameters. The reason for this recommendation is that RMSE, being in units of the original dependent variable, is most straightforward for a reader to intuitively grasp and evaluate.

Parameters expressed in intuitive units

In the same spirit, the results of applying a mathematical model are best understood if the parameters themselves are expressed natural and well-defined units. Parameter units such as prob-

[14] By which is meant that equations for the best-fitting parameter values can be generated, e.g., for a linear-regression model, slope = $(n\Sigma XY - \Sigma X \Sigma X)/[n\Sigma X^2 - (\Sigma X)^2]$.

ability correct raised to the 1.6 power, for instance are not intuitively appealing, while parameter units such as time (e.g., milliseconds) are much more intuitively appealing. When parameters are defined in natural units, results of experiments can be conveyed in terms of effects of independent variables on parameter values, which is considerably simpler than trying to describe the data in terms of, say, a set of complex interactions. A simple example of such a model is Sternberg's (e.g., 1967) short-term scanning model in which two parameters—the scanning time per item and a "time for everything else" parameter—are both defined in units of time (milliseconds). Given the validity of the model, the results of any given experimental Sternberg-task condition can be described by one number—the slope of the search function, which is an estimate of the scanning time per item. Different conditions, e.g., degraded versus undegraded target-item conditions, can then be described simply in terms of the degree to which scanning time differs over the different conditions.

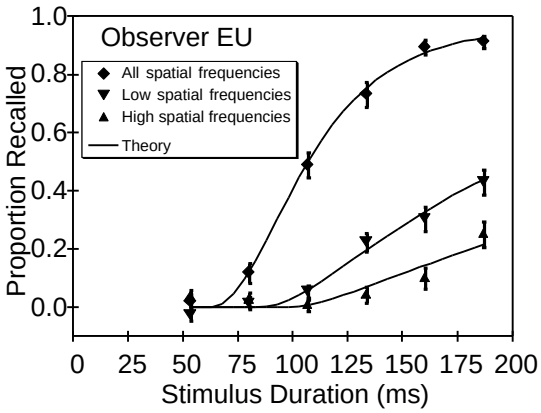

Figure 13. Unpublished data from Harley and Loftus showing digit recall performance as a function of stimulus duration and the nature of spatial filtering. Smooth lines through the data points represent theoretical predictions based on the best-fitting parameter values. The error bars are standard errors.

Model fitting and hypothesis testing

Thus far, I have treated model fitting and hypothesis testing as separate enterprises. At their core, however, they are the same thing. In both instances, a model is proposed, and the data are treated in such a way as to evaluate the plausibility of the data given that the model is correct.

The difference between model fitting and hypothesis testing is one of tradition, not of substance. In a hypothesis-testing procedure, the null hypothesis is almost invariably that some set of population means are all equal to one another. However, such a characterization of the null hypothesis is not necessary; as suggested in the above section on planned comparisons, *any* single-degree-of-freedom hypothesis is a valid null hypothesis. Thus, the best-fitting set of parameter values issuing from the fit of a mathematical model to a data set can be characterized as a null hypothesis and tested with the standard hypothesis-testing machinery. Note that there are two other departures from tradition involved in this process. First, the investigator's goal is typically to *accept* the null hypothesis (i.e., to confirm the model) rather than to reject the null hypothesis and second, reliance on the Linear Model is deemphasized considerably—the mathematical model being tested could be linear, but it often is not.

Display of Data Fits

My final set of comments about mathematical models revolve around displays of data fits. As just indicated, many experimental results can most parsimoniously be expressed as effects of independent variables on parameter values. This technique works best when the model under consideration is well-tested and accepted as an accurate description of the experimental paradigm under investigation. With this kind of mature model, the validity of the model is not under consideration; rather the model is being used as a tool to investigate something else (in the Sternberg example above, to investigate the effect of stimulus degradation on search slope—where "slope" is preaccepted as a meaningful entity within the context of the Sternberg model).

With less developed models, however, a central issue is often whether (or the degree to which) the model is adequately fit by the data to begin with. In this case, the main result to be presented is the model fit itself. As noted, the most straightforward way of doing this is with a single number, the RMSE. However, a graphical display of the model fit is also critical in order that systematic failures of the model can be highlighted. How this is done depends on the relation between what the model predicts and the dependent variable measured in the experiment.

Quantitative Predictions

When the model is sufficiently precise that it predicts a specific value of the dependent variable for each experimental condition, the fit can be presented as a standard graph of the observed data plus predicted data. As an example, consider data from my laboratory generated by an experiment in which 4-digit strings were presented at varying durations for immediate recall. The strings were either spatially filtered so as to leave only low spatial frequencies, only high spatial frequencies,

or were presented normally, i.e., including all spatial frequencies. A mathematical model described by Olds and Engel (1998), which predicted an exact value of the dependent variable (proportion recalled) for each condition was fit to the data. The data and model fit are shown in Figure 13. I wish to emphasize several aspects of the data presentation.

1. The data are presented as symbols only (diamonds and triangles) while the model fits are presented as symbol-less lines.

2. The model predictions are "complete." By this I mean that the theoretical lines include predicted fits not just for the discrete durations selected for the experiment, but continuously over the selected range. This means that the predicted curves are smooth, and the predictions of the theory are clearer than would be the case if only the predictions corresponding to the experimental durations were shown.

3. Finally, at the risk of sounding overly compulsive, it is mnemonically wise, and aesthetically elegant, to select, if possible, data symbols that are somehow naturally associated with the conditions. In the case of Figure 13, for example, downward-pointing triangles represent low spatial frequencies, upward-pointing triangles represent high spatial frequencies, and diamonds (i.e., the superimposition of downward- and upward-pointing triangles) represent all spatial frequencies.

Monotonic Predictions

Sometimes a mathematical model predicts some quantity that may be assumed only monotonically related to the dependent variable measured in the experiment. For example, Loftus & Irwin (1998) developed a theory of missing-dot, temporal-integration performance (e.g., Di Lollo, 1980). In this paradigm, 24 dots are briefly presented to a subject as a 5x5 dot array with one dot missing. The subject's task is to identify the position of the missing dot, and missing-dot detection probability is the dependent variable. The dots are presented in two temporal halves: During half 1, a random 12 of the 24 dots are presented, while the remaining 12 dots are presented during half 2. This means that in order to perform the missing-dot detection task, the subject must integrate the spatial information corresponding to the two dot-array halves over time. For purposes of the present discussion, the duration of half 2 was short (20 ms) while the duration of half 1 varied from 20 to 100 ms in 20-ms steps, and the duration of the ISI separating the end of half 1 from the start of half 2 varied from -20 to 60 ms in 20-ms steps.

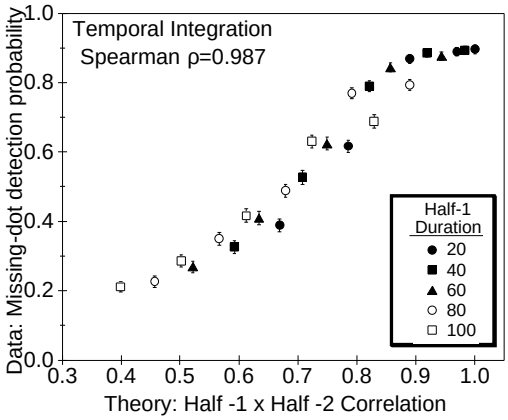

Figure 14. Obtained data as a function of a predicted theoretical construct for a missing-dot, temporal-integration experiment. The different curve symbols correspond to different values of half-1 duration. The 5 data points within each half-1 duration correspond to different ISI values (data from Loftus & Irwin, 1998). The error bars are standard errors.

Central to Loftus and Irwin's theory was the proposition that a visual stimulus triggers an internal *sensory-response function* that rises over time beginning at stimulus onset and falls, eventually back to zero, following stimulus offset (see also Busey & Loftus, 1994; Loftus and McLean, 1999). In the missing-dot paradigm, each stimulus half produces one such sensory-response function, and performance is determined by—that is, is a monotonic function of—the correlation over time between the two sensory-response functions (as suggested by Dixon & Di Lollo, 1994). The theory can specify the magnitude of this correlation for any stimulus condition, but it does not specify the nature of the monotonic function that relates missing-dot detection probability to correlation magnitude.

In this kind of situation, it is not possible to fit the theory using the techniques listed above because they require that the theory predict the actual dependent variable, not just something presumed to be monotonically related to the dependent variable. A straightforward alternative is to use as a fit criterion the rank-order correlation (Spearman ρ) over conditions between the data and the theory. The fit may then be represented as the data-theory scatterplot which, if the data fit the theory perfectly, would be monotonic.

Figure 14 provides an example, using the paradigm and theory that I have just described. The predicted data points are generated by the parameter values corresponding to the highest data x theory Spearman ρ (ρ = 0.987) found by the search procedure. The scatterplot shows data (mean proportion of correctly detected missing-dot

positions) plotted against theory (predicted correlation) across the 25 half-1 duration x ISI conditions. Within the scatterplot, different half-1 durations are represented by different curve symbols, while within each half-1 duration, increasing predicted and observed values correspond to decreasing ISI's.

Presenting the fit as a scatterplot confers at least two benefits. First, the shape of the scatterplot (which is ogival in this example) constitutes an empirical estimate of the actual monotonic function relating the dependent variable to the theoretical construct, thereby providing clues about the mechanism that relates the dependent variable to the theoretical construct. Second, the scatterplot underscores systematic discrepancies in the data fit. In this example, it appears that performance in the long half-1 duration conditions (e.g., the 100-ms half-1 duration conditions, represented by the open squares) are observed to be systematically higher than they are predicted to be compared to short half-1 duration conditions (e.g., the 20-ms half-1 duration conditions represented by the solid circles) thereby pinpointing a specific deficit in the theory.

Equivalence Techniques to Investigate Interactions

In the large majority of psychological experimentation, an investigator sets the levels of some independent variable and measures the resulting values of the dependent variable. Of interest then is how changes in the independent variable lead to changes in the dependent variable. Moreover, a sizeable number of experiments are primarily concerned not with main effects, but with how one or more independent variables *interact* in their effects on the dependent variable.

As numerous writers have pointed out (e.g., Bogartz, 1976; Loftus, 1978), many conclusions resting on interactions have strong limitations, the most severe of which is that nonordinal (i.e., noncrossover) interactions lack generality both with respect to other performance measures that are nonlinearly related to one actually measured (e.g., an interaction in a memory experiment observed in terms of probability correct cannot necessarily be generalized to d') and also with respect to underlying theoretical constructs that are nonlinearly related to the performance measure (e.g., an interaction observed in terms of probability correct cannot necessarily be generalized to some generically defined "memory strength").

To circumvent these difficulties one can turn to *equivalence techniques*, which are a set of theoretical/methodological procedures by which one determines the rules under which different combinations of independent variables lead to equivalent states of some inferred internal psychological state. Equivalence techniques have roots in classical statistics (e.g., Hays, 1973) and conjoint measurement (e.g., Krantz, Luce, Suppes & Tversky, 1971; Krantz & Tversky, 1971; Tversky & Russo, 1969).

Equivalence techniques are common in vision science. Perhaps the best illustration of how such techniques are used to understand the workings of the visual system is the classic *color-matching experiment*, wherein an observer adjusts some additive combination of primary colors such that it matches a monochromatic test color (e.g., Wright, 1929; 1946). The resulting two stimuli—the combination of primaries and the monochrome stimulus—constitute *color metamers*, which, though entirely different physically, are equivalent psychologically in a fundamental way: They entail equal quantum catches in the three classes of cone photoreceptors. The original success of the color-matching experiment constituted the empirical foundation of the trichromacy theory of color vision, and versions of the color-matching experiment have been used more recently to refine and modify the theory (e.g., Wandell, 1982).

As will be discussed in the next two subsections, there are two ways in which equivalence techniques can be used in virtually any area of psychology. First, relatively weak hypotheses about effects of certain variables can be studied using *state-trace analysis*. Second, stronger hypotheses make specific, unique, and testable predictions about the specific quantitative rules by which multiple independent variables combine to produce equivalent values of the dependent variable.

State-Trace Analysis

State-trace analysis was introduced by Bamber (1979) as a means of investigating relations among independent variables, dependent variables, and hypothesized internal dimensions. In particular, state-trace analysis can be used to answer two related questions. First, is the assumption of a single internal dimension sufficient to account for observed relations among multiple independent variables and multiple dependent variables? Second, if more than one dimension is necessary, what are the characteristics of the multiple dimensions; i.e., how are they affected by the independent variables and how do they influence the dependent variables?[15]

[15] Examples of state-trace analysis are rare in most of psychology. In addition to examples provided by Bamber (1979) and Loftus et al. (2000), see Loftus

To illustrate the use of state-trace analysis, consider a face-recognition investigation described by Busey, Tunicliff, Loftus, & Loftus (2000). The experimental paradigm entailed an initial study phase wherein a series of face pictures was sequentially presented, followed by a yes-no recognition test phase in which two dependent variables—accuracy (hit probability) and confidence (on a four-point scale)—were measured. Of principal interest was whether accuracy and confidence were simply two measures of the same internal state which, for mnemonic convenience, might be termed "strength". The experiment entailed two independent variables that were manipulated during the study phase. First, exposure duration varied, and second, each studied face was followed by a 15-sec period during which visual rehearsal of the just-seen face was either required or prohibited. The main results were, unsurprisingly, that both accuracy and confidence increased with increasing exposure duration and with rehearsal compared to no rehearsal. That is, qualitatively both accuracy and confidence were affected in the same way by the two independent variables, thereby suggesting, in the tradition of "dissociation techniques," that they were simply two measures of the same internal state.

However, the use of state-trace analysis allowed a much more precise answer to the question. More specifically, the proposition that any two dependent variables—accuracy and confidence in this instance—are measures of the same internal state can be couched in the form of a hypothesis, called the "*single-dimension model*," which is: "there exists a single internal dimension (call it "strength") whose value is jointly determined by duration and rehearsal, and which, in turn, determines the values of both confidence and accuracy. The form of this model is shown at the top of Figure 15. Pitted against this single-dimensional model is some form of multi-dimensional model, according to which the two dependent variables are determined at least in part by different internal dimensions. While a single-dimensional model (akin to a standard null hypothesis) is unique, there are an infinite number of possible multi-dimensional models (akin to there being an infinite number of alternative hypotheses). One reasonable multi-dimensional model is shown at the bottom of Figure 15. Here, a second dimension, termed "certainty" is affected by rehearsal but not duration, and affects confidence, but not accuracy.

The key prediction of the single-dimensional hypothesis rests on the logic that any two conditions—a long-duration no-rehearsal condition, and a shorter-duration, rehearsal condition—that produce equal accuracy must have done so because they produced the same strength values. Thus—and here is the key prediction—because confidence is also determined by strength, these same two conditions must also produce equal confidence values.

To evaluate this prediction, one constructs *state-trace plots*, which are scatterplots of one dependent variable plotted against the other (accuracy plotted as a function of confidence, in this instance) over the experimental conditions defined by the combination of the two independent variables—in this case, conditions defined by the duration x rehearsal combinations. The prediction then translates to: The curve traced out by the rehearsal conditions must overlie the curve traced out by the no-rehearsal conditions.

It should be noted, incidentally, that the success of state-trace analysis does not require that one be lucky enough to find pairs of duration x rehearsal conditions that happen to produce identical performance. The formal rationale for this assertion is described in Bamber (1979). Essentially, one assumes that the measured points are samples from an underlying continuous function whose form can be estimated by "connecting the dots" in the state-trace plot.

Figure 16 shows predictions from the single-dimensional model (top panels) and from the multi-dimensional model (bottom panels) of Figure 15. In each panel, circles correspond to the rehearsal conditions while triangles correspond to the no-rehearsal conditions. The 5 instances of each curve symbol correspond to 5 exposure durations. For each of the two models, the two left-hand panels ("Accuracy" and "Confidence") present the data in the manner in which such data are normally presented: The dependent variable is plotted as a function of the independent variables (duration along the abscissa and rehearsal as the curve parameter in this example). Based on these standard data, there is nothing very obvious that distinguishes the predictions of the two models.

However, the state-trace plots shown as the rightmost panels ("C-A Scatterplot") distinguish strongly between the two models. As described above, the single-dimensional model predicts that the two scatterplots corresponding to the two rehearsal levels fall atop one another. However, the multi-dimensional model predicts that the two scatterplots are distinguishable in some manner

---

and Irwin (1998) who used such analyses to address the question: "Are visible persistence and iconic memory just two names for the same internal process?" Palmer (e.g., 1986a; 1986b) has used related (and formally identical) equivalence techniques to examine numerous issues in attention and perception.

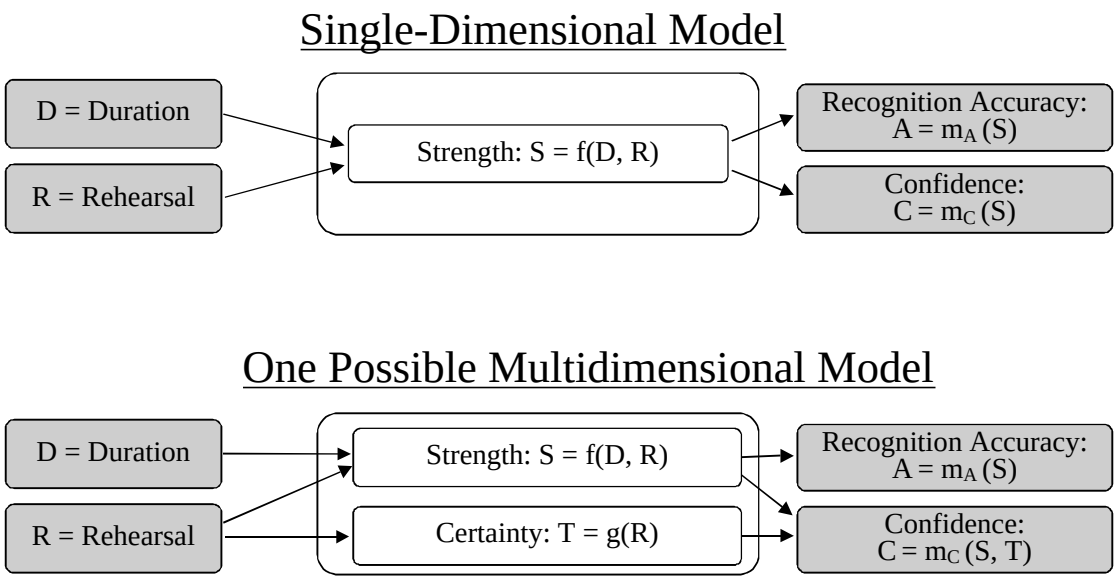

Figure 15. Two models on the relations between two independent variables and two dependent variables in a face-recognition experiment (reported by Busey et al. 2000). The left-hand shaded rounded rectangles represent independent variables, while the right-hand shaded rounded rectangles represent dependent variables. The middle unshaded rounded rectangles represent unidimensional theoretical constructs.

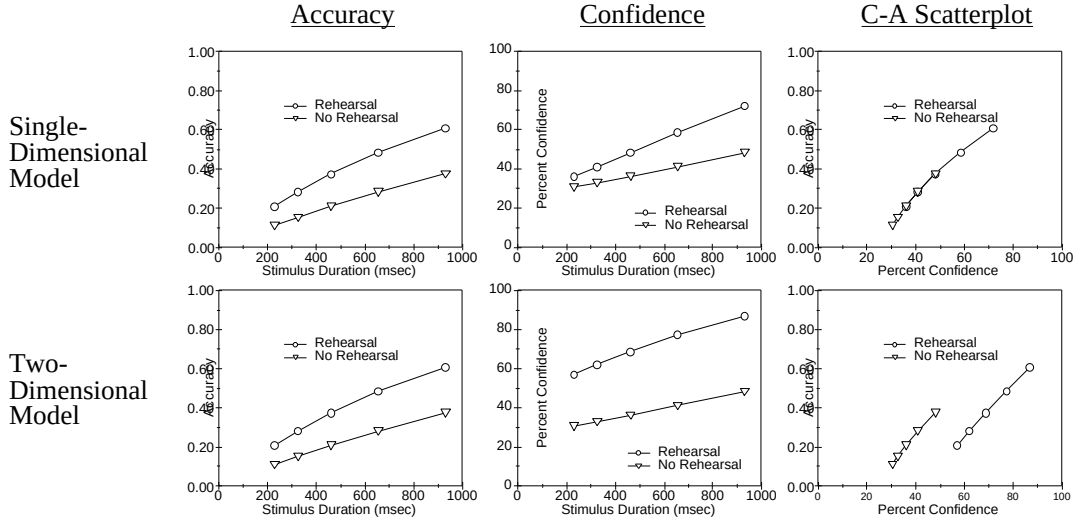

Figure 16. Theoretical predictions from the models shown in Figure 15. The left and middle panels show "standard" data presentation: the dependent variable plotted as functions of the independent variables. The right panels show state-trace plots which are scatterplots of one dependent variable plotted against the other. With the state-trace plots, the distinct predictions of the two models are considerably more apparent than they are in the standard plots.

that depends on the exact construction of the multi-dimensional model. In the multi-dimensional model of Figure 15, a second internal dimension, "certainty" is increased by rehearsal but not duration, and increased certainty increases confidence but not accuracy. Therefore, according to this particular multi-dimensional model, two conditions that produce the same accuracy values

must have done so because they produced the same strength value. However, comparing two conditions that produce the same strength values, the rehearsal condition will produce greater confidence than the no-rehearsal condition, because certainty is greater in the rehearsal condition than in the no-rehearsal condition. Therefore, as is evident in the prediction (Figure 16, bottom-right panel) this particular multi-dimensional model predicts the rehearsal curve to be displaced to the right of the no-rehearsal curve.

In sum, state-trace analysis has two virtues. First, it allows one to test any form of single-dimensional model, which is generally a strong test of the common question in psychology: "Are two dependent variables, Y and Y' simply two measures of the same internal state?" Second, given that one rejects a single-dimensional model, the resulting form of the state-trace plots, provides strong clues as to the nature of the implied multi-dimensional model. To briefly illustrate, Busey et al. (1990) actually investigated two kinds of confidence: Prospective confidence, given at the time of study, and retrospective confidence, given at the time of test. They determined that a single-dimensional model was appropriate to describe retrospective confidence, but that a multi-dimensional model of the sort depicted at the bottom of Figure 15 was necessary to describe prospective confidence.

Additive and Multiplicative Effects

As just described, state-trace analysis deals with the qualitative question: Do multiple independent variables affect the same internal memory dimension which then determines performance in the relevant memory tasks? An investigator can also use equivalence techniques to unveil stronger quantitative rules by which independent variables combine to produce a value on the internal dimension. To illustrate such rules, I will use two examples in which memory is measured as a function of the exposure duration of the to-be-remembered stimulus (as in the Busey et al., 2000 experiment described in the last section). In this kind of experimental paradigm, define a *performance curve* as a curve that relates memory performance to exposure duration (as, for example, in Figure 16, four left panels.) Define a *focal variable* as some variable under consideration that is factorially combined with exposure duration (e.g., rehearsal in the Busey et al. experiment). The equation relating performance curves for two levels of the focal variable is:

$$p [i, d] = p[j, f(d)] \qquad \text{(Eq. 21)}$$

where p[i, d] and p[j, f(d)] denote performance for levels i and j of the focal variable, d and f(d) are durations, and f is a monotonic function. Again in the spirit of equivalence, it is important to realize that Equation 21 describes duration relations that produce equal performance for different focal-variable levels.

Of theoretical interest in a given situation is the nature of the function f(d) on the right side of Equation 21. Different f(d)'s are implied by different hypotheses about the focal variable's effect. I illustrate with two common hypotheses. The first is that the focal variable's effect is *additive*, i.e., that f(d) = d+k, which means that,

$$p (i, d) = p(j, d+k) \qquad \text{(Eq. 22)}$$

Here, k is a constant in units of time. The interpretation of an additive effect is that being in level i of the focal variable is equivalent to having an additional k ms of stimulus duration compared to being in level j. As shown in Figure 17A, stimulus masked/not masked exemplifies an additive focal variable with k = 100 ms—which is the basis for the claim made by Loftus, Johnson & Shimamura (1985) that "an icon is worth 100 ms."

The second hypothesis is that the focal variable's effect is *multiplicative*, i.e., that f(d) = cd, which means that,

$$p(i, d) = p(j, cd) \qquad \text{(Eq. 23)}$$

Here, c is a dimensionless constant. The interpretation of a multiplicative effect is that being in level j of the focal variable slows down processing by a factor of c, compared to being in level i. As shown in Figure 17B, stimulus luminance exemplifies a multiplicative focal variable with c = 2 (see Loftus, 1985b; Sperling, 1986; as shown by Loftus & Ruthruff, 1994, the same is true when contrast is the focal variable).

Figure 17 illustrates three important facets of using equivalence techniques. First, testing various hypotheses (e.g., that the effect of some focal variable is additive or multiplicative) involves *horizontally comparing* performance curves because, as indicated in Equations 16-18, the critical comparisons are of the durations d and f(d) required to achieve equal performance for different focal-variable levels, i and j. Second, an additive hypothesis predicts that performance curves be *horizontally parallel* as in Figure 17A, whereas a multiplicative hypothesis predicts that performance curves be *constant-ratio diverging* as in Figure 17B. Third, Figure 17C demonstrates that a multiplicative hypothesis can be conveniently tested by plotting performance on a log-duration scale

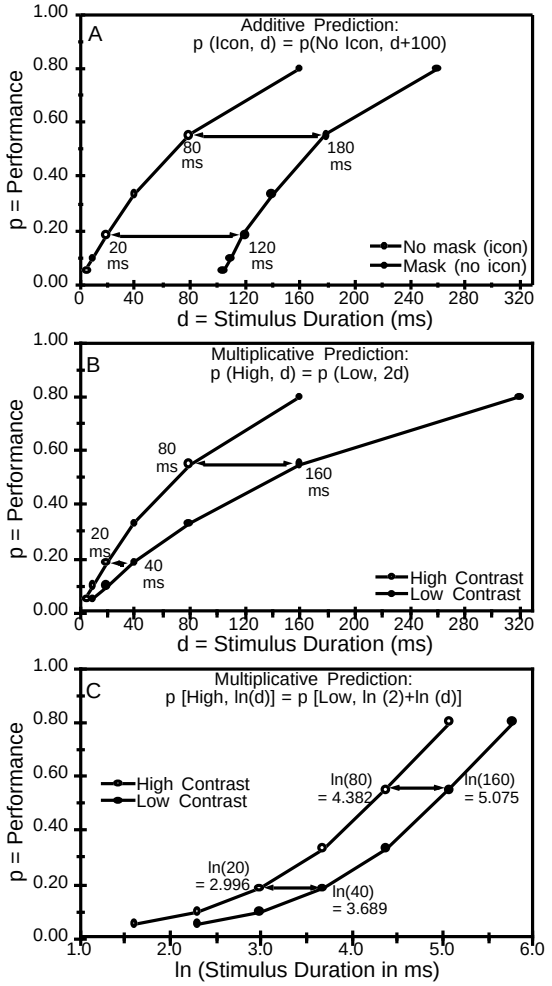

Figure 17. Panel A: Performance as a function of stimulus duration and stimulus masked/not mask. The horizontally parallel curves (after Loftus et al., 1995) reflect an additive effect of masking: The iconic image eliminated by the mask is worth 100 ms of additional physical exposure duration. Panel B: Stimulus contrast (high, low) replaces stimulus masked/not masked. Here a multiplicative result occurs: The exposure duration required to achieve a constant performance level is greater (by a factor of 2) for the low-contrast compared to the high-contrast condition (after Loftus, 1985). Panel C: The Panel-B multiplicative relation plotted on a log-duration axis produces easy-to-test horizontally parallel curves rather than the difficult-to-test constant-ratio diverging curves of Panel B.

instead of a linear-duration scale. When d is on a log scale, Equation 23 becomes,

$$p[i, \ln(d)] = p[j, \ln(c) + \ln(d)]$$

and performance curves are again horizontally parallel, separated by a constant of ln (c), which can then, of course, be exponentiated to recover c.

As I asserted earlier, equivalence techniques represent scale-independent means of identifying the fundamental nature of interactions among variables. Equivalence techniques allow conclusions that are more generalizable and robust than are conclusions based on most traditional statistical interactions. Because performance curves are compared horizontally, any conclusion issuing from the comparison (e.g., that the curves are or are not horizontally parallel on a linear or on a log-duration scale) is invariant over all monotonic transforms of the performance measure; this is because any set of points that are equal in one scale must also be equal in any monotonically related scale. Therefore, conclusions issuing from equivalence techniques apply not only to the particular dependent variable being measured (e.g., proportion correct) but also to any theoretical construct that is assumed to be monotonically related to the dependent variable (e.g., "memory strength"). Such conclusions also apply, *mutatis mutandis*, to any dependent variable that is monotonically related to the dependent variable being measured (e.g., to d' if the measured variable is proportion correct).

CONCLUSIONS

It is worth concluding by briefly reiterating the sentiments expressed at the outset of this chapter. Lurking within a typical data set is often a wealth of fascinating information that can be summoned forth if sufficiently clever detective techniques are used. As has been argued in many places (see particularly, Loftus, 1996; Schmidt, 1996) there are, at present, many standard data-analysis techniques that not only are ill-crafted for eliciting such information, but actively bias the investigator against finding anything interesting or non-obvious from the data. It is my hope that some of the less common techniques described in this chapter—and other related techniques that the reader is left to devise on his or her own—will provide some assistance in coping with the vast sea of psychological data that our present technology currently produces for us.

References

Abelson, R.P. (1995). *Statistics as Principled Argument*. Hillsdale NJ: Lawrence Erlbaum Associates.

Bogartz, R.S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child psychology, 22*, 178 183.

Busey, T.A. & Loftus, G.R. (1994). Sensory and cognitive components of visual information acquisition. *Psychological Review, 101*, 446-469.

Busey, T.A., Tunnicliff, J., Loftus, G.R. & Loftus, E.F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26-48,

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378-399.

Carver, R.P. (1993). The case against statistical significance testing, rivisited. *Journal of Experimental Education, 61,* 287-292.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Di Lollo, V. (1980). Temporal Integration in Visual Memory. *Journal of Experimental psychology: General, 109,* 75 97.

Dixon, P. & Di Lollo, V. (1994). Beyond visible persistence: An alternative account of temporal integration and segregation in visual processing. *Cognitive psychology* (in press).

Frick, R.W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132-138

Gigerenzer, G. , Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L.. (1989)*The Empire of chance: How probability changed science and everyday life.* Cambridge England: Cambridge University Press

Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69,* 54-61.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 83,* 1-20.

Guilford, J.P. (1942). *Fundamental Statistics in psychology and Education.* First Edition, 1942; third edition, 1956. New York: McGraw-Hill.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (eds.) (1997). *What if there were no significance tests?* Mahwah, JH: Lawrence Erlbaum Associates.

Hays, W. (1973). Statistics for the Social Sciences (second edition). New York: Holt.

Jackson, J.M. & Latane, B. (1981). All alone in front of all those people: Stage fright as a function of number and type of co-performers and audience. *Journal of Personality and Social psychology, 40,* 73-85.

Jones, L.V. (1955). Statistics and research design. *Annual Review of psychology, 6,* 405-430. Stanford: Annual Reviews, Inc.

Krantz, D.H. & Tversky, A. (1971). Conjoint measurement analysis of composition rules in psychology. *Psychological Review, 78,* 151 169.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 44, 1372-1381.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement*. New York: Academic Press.

Loftus, G.R. (1978). On interpretation of interactions. *Memory and Cognition, 6,* 312 319.

Loftus, G.R. (1985). Consistency and confoundings: Reply to Slamecka. *Journal of Experimental psychology: Learning, Memory and Cognition, 11,* 817-820. (b)

Loftus, G.R. (1993). Editorial Comment. *Memory & Cognition, 21,* 1-3. (b)

Loftus, G.R. (1985). Evaluating forgetting curves. *Journal of Experimental psychology: Learning, Memory and Cognition*, 11, 396-405. (a)

Loftus, G.R. (1985). Picture perception: Effects of luminance level on available information and information-extraction rate. *Journal of Experimental psychology: General, 114,* 342-356 (b)

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary psychology. 36,* 102-105.

Loftus, G.R. (1993). A picture is worth a thousand *p*-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers, 25,* 250-256. (a)

Loftus, G.R. (1995). Data analysis as insight. Behavior Research Methods, Instrumentation and Computers, 27, 57-59.

Loftus, G.R. (1996). psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 161-171.

Loftus, G.R. & Bamber, D. (1990) Weak models, strong models, unidimensional models, and psychological time. *Journal of Experimental psychology: Learning, Memory, and Cognition, 16,* 916-926.

Loftus, G.R. & Irwin, D.E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive psychology, 35,* 135-199.

Loftus, G.R., Johnson, C.A., & Shimamura, A.P. (1985). How much is an icon worth? *Journal of Experimental psychology: Human Perception and Performance, 11,* 1-13.

Loftus, G.R., Johnson, C.A., & Shimamura, A.P. (1985). How much is an icon worth? *Journal of*

Experimental psychology: Human Perception and Performance, 11, 1-13.

Loftus, G.R. & Masson, M.E.J. (1994) Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review, 1,* 476-490.

Loftus, G.R. & McLean, J.E. (1999). A front end to a theory of picture recognition. *Psychonomic Bulletin & Review,* 6, 394-411.

Loftus, G.R. & Ruthruff, E.R. (1994). A theory of visual information acquisition and visual memory with special application to intensity-duration tradeoffs. *Journal of Experimental Psychology: Human Perception and Performance,* 20, 33-50.

Lykken, D.T. (1968). Statistical significance in psychological research. *Pschological Bulletin, 70*, 131-139.

Maltz, M.D, (1994). Deviating from the mean: The declining significance of significance. Journal of research in crime and delinquency, 31, 434-463.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P.E. (1978). Theortical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow process of soft psychology. *Journal of Consulting and Clinical psychology, 46,* 806-834.

Meehl, P.E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, Monograph Supplement 1-V66*.

Melton, A.W. (1962). Editorial. *Journal of Experimental psychology, 64*, 553 557.

Olds, E.S. & Engel, S.A. (1998). Linearity across spatial frequency in object recognition. *Vision Research, 38*, 2109 2118.

Palmer, J.C. (1986). Mechanisms of displacement discrimination with and without perceived movement. *Journal of Experimental psychology: Human Perception and Performance, 12*, 411 421. (a)

Palmer, J.C. (1986). Mechanisms of displacement discrimination with a visual reference. *Vision Research, 26*, 1939 1947. (b)

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Schmidt, Frank (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

Schmidt, Frank & Hunter, J. (1997). Eight false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow & S. Mulaik (Eds.), What if there were no significance testing. Hillsdale, NJ: Erlbaum.

Slamecka, N.J. & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. Journal of Experimental psychology: Learning, Memory, and Cognition, 9, 384 397.

Slamecka, N.J. (1985). On comparing rates of forgetting. Journal of Experimental psychology: Learning, Memory & Cognition, 11, 812 816.

Smith, A.F. & Prentis, D.A. (1993). Graphical Data Analysis. In G. Kerens, and C. Lewis (Eds.) A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues, Hillsdale NJ: Erlbaum.

Sperling, G. (1986). A signal to noise theory of the effects of luminance on picture memory: Commentary on Loftus. Journal of Experimental psychology: General, 115, 189 192.

Sternberg, S. (1967). Two operations in character recognition: Some evidence from reaction time measurements. Perception and Psychophysics 2, 43 53.

Tufte, E.R. (1983). The visual display of quantitative information. Cheshire, CT: Graphics Press.

Tufte, E.R. (1990). Envisioning information. Cheshire, CT: Graphics Press.

Tukey, J.W. (1977) Exploratory data analysis. Reading, MA: Addison Wesley.

Tversky, A., & Russo, J.E. (1969). Substitutibility and similarity in binary choices. Journal of Mathematical psychology, 6, 1 12.

Tyler, R.W. (1935). What is statistical significance? Educational Research Bulletin, 10, 115-118, 142.

Wainer, H. & Thissen, D. Graphical Data Analysis (1993). In G. Kerens, and C. Lewis (Eds.) A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues, Hillsdale NJ: Erlbaum.

Wandell, B. A. (1982) Measurements of small color differences. Psychological Review, 89, 281-302.

Wright, W.D. (1929). A re determination of the trichromatic coefficients of the spectral colours. Transactions of the Optical Society of America, 30, 141 164.

Wright, W.D. (1946). Researches on normal and defective colour vision. London: Henry Kimpton.