Psychology Will Be a Much Better Science When We Change the Way We Analyze Data

Geoffrey R. Loftus[1]

Department of Psychology, University of Washington, Seattle, Washington

In 1964, I entered the field of psychology because I believed that within it dwelt some of the most fundamental and challenging problems of the extant sciences. Who could not be intrigued, for example, by the relation between consciousness and behavior, or the rules guiding interactions in social situations, or the processes that underlie development from infancy to maturity? Today, in 1996, my fascination with these problems is undiminished. But I have developed a certain angst over the intervening 30-something years—a constant, nagging feeling that our field spends a lot of time spinning its wheels without really making much progress. This problem shows up in obvious ways—for instance, in the regularity with which findings seem not to replicate. It also shows up in subtler ways—for instance, one does not often hear psychologists saying, "Well, this problem is solved now; let's move on to the next one" (as, e.g., Johannes Kepler must have said more than three centuries ago, after he had cracked the problem of describing planetary motion).

I have come to believe that at least part of this problem revolves around our tools—particularly the tools that we use in the critical domains of data analysis and data interpretation. What we do, I sometimes feel, is akin to trying to build a violin using a stone mallet and a chain saw. The tool-to-task fit is not very good, and, as a result, we wind up building a lot of poor-quality violins.

My purpose here is to elaborate on these issues. In what follows, I summarize our major data-analysis and data-interpretation tools and describe what I believe to be amiss with them. I then offer some alternative techniques for extracting more insight and understanding from a data set.

Recommended Reading

Abelson, R.P. (1995). (See References)
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L, Beatty, J., & Kruger, L. (1989). (See References)
Loftus, G.R. (1991). (See References)
Rosnow, R.L., & Rosenthal, R. (1989). (See References)
Schmidt, F. (1996). (See References)

THE UNIVERSALITY OF NULL-HYPOTHESIS SIGNIFICANCE TESTING

The vast bulk of data analysis and data interpretation in the social and behavioral sciences is carried out using a set of techniques collectively known as *null-hypothesis significance testing* (NHST). The logic of NHST goes as follows.

1. An investigator begins with the hypothesis that some independent variable will have an effect on some dependent variable. At its most general level, the hypothesis can be expressed as

$$\text{not } (\mu_1 = \mu_2 = \ldots = \mu_J), \quad (1)$$

where $\mu_1, \ldots, \mu_J$ are the means[2] of J population distributions of the dependent variable that correspond to J levels (i.e., values) of the independent variable. Equation 1 is generically referred to as an *alternative hypothesis*, or H_1.

2. Next, the investigator conducts an experiment in which J random samples of the dependent variable are obtained—one sample for each level of the independent variable. This experiment yields the observed sample means $M_1, \ldots, M_J$, which are estimates of the population means $\mu_1, \ldots, \mu_J$. Generally, it is not true that $M_1 = M_2 = \ldots = M_J$; that is, there will always be some differences among the sample means. What needs to be determined is whether the observed differences among the sample means are due only to random errors in measurement or whether they are due, at least in part, to corresponding differences among the population means. Informally, the investigator needs to provide a convincing argument that the observed effect is "real."

3. To this end, the investigator endeavors to compute the probability (known as p[3]) of observing differences among the Ms as great as those that were actually observed given that, in fact, the J population means are all equal, that is, given that

$$\mu_1 = \mu_2 = \ldots = \mu_J. \quad (2)$$

This hypothesis that the population means are equal is referred to as the *null hypothesis*, or H_0.

4. Based on a comparison of the computed value of p with a criterion value of p known as α (α is usually set at .05), the investigator makes a binary decision. If p is less than α, then the investigator makes a strong decision to reject H_0 in favor of H_1 (a decision that is usually phrased as "the observed effect is statistically significant"). If p is greater than α, then the investigator makes a weak decision to fail to reject H_0.

   Two types of errors can be committed in this setting. One commits a *Type I error* when one incorrectly rejects a true null hypothesis. If the null hypothesis is true, then the probability of a Type I error is, by definition, equal to α. One commits a *Type II error* when one incorrectly fails to reject a false null hypothesis. If the null hypothesis is false, a Type II error is committed with probability β. In general, we do not know the value of β because we have no information, and no assumptions, regarding the values of the actual μ's, given that the null hypothesis is false. Statistical power is defined to be $(1 - \beta)$, which is interpreted as the probability of correctly rejecting a false null hypothesis. Because β is not generally known, neither is power.

5. Finally, based on a series of such decisions—that is, rejecting or failing to reject a series of null hypotheses—the investigator tries to make sense of the data set, no matter how complex it might be.

Although variants of this procedure constitute the primary means of making conclusions from the vast majority of psychology experiments, I do not believe that it is a fruitful way of interpreting data or understanding psychological phenomena. On the contrary, I believe that reliance on NHST has channeled our field into a series of methodological cul-de-sacs, and it has been my observation over the years (particularly over my 4 years as editor of *Memory & Cognition*) that conclusions made entirely or even primarily based on NHST are at best severely limited, and at worst highly misleading. In the following section, I articulate the reasons for these beliefs.

I am by no means the first person to issue such charges. Periodically, a book or an article decrying the enormous reliance we place on NHST will appear.[4] Sadly, however, although such airings of the issues occasionally attract attention, they have not (up until now, anyway) impelled widespread action.[5] They have been carefully crafted and put forth for consideration, only to just kind of dissolve away in the vast acid bath of existing methodological orthodoxy.

SIX THINGS TO NOT LIKE ABOUT NHST

In this section, I articulate six major problems with NHST. As I have indicated, they have been described before. But they bear repeating, and it is useful to consider them in concert.

The Usual Impossibility of a Typical Null Hypothesis

NHST usually revolves around the testing of a null hypothesis that could not really be true to begin with (see Meehl, 1967, pp. 108–110, for a careful articulation of this issue). For example, suppose an investigator presented subjects with digit strings on a computer screen to study the effects of stimulus duration on the subsequent recall of the digit strings. Each digit string might be shown for one of five exposure durations ranging from, say, 10 ms to 100 ms, and the investigator would measure the proportion of digits correctly recalled in each condition. In the usual hypothesis-testing framework, the investigator would establish the following null hypothesis:

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5, \quad (3)$$

where the μ's refer to the population means of the percentage of digits recalled for each of the five exposure durations. Note that the logic of NHST demands that the equal signs in Equation 3 mean "equal to an infinite number of decimal places." If one weakens this requirement such that the equal signs mean "pretty much equal," then one must add additional assumptions specifying what is meant by "pretty much." Although the mathematical machinery for doing this has been worked out (e.g., Hays, 1973), it is rarely (if ever) implemented in practice.

This null hypothesis of identical population means cannot be literally correct. As Meehl (1967) has pointed out,

> 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. (p. 109)

Alternatively, the μ'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.[6]

Accordingly, differences in exposure duration must lead to performance differences even if such differences are small, and the relevant question is not really whether there are any differences among the population means. Rather, the relevant questions are, How big are the differences? Are they big enough for the investigator to care about, and, if so, what pattern do they form? In short, testing the null hypothesis of Equation 2 cannot provide new information. All it can do is indicate whether there is enough statistical power to detect whatever differences among the population means must be there to begin with. As I have noted elsewhere (Loftus, 1995), rejecting a typical null hypothesis is like rejecting the proposition that the moon is made of green cheese. The appropriate response would be "Well, yes, okay . . . but so what?"

"Significance" Versus the Underlying Pattern of Population Means

A finding of statistical significance only constitutes evidence (and vague evidence at that, as we shall see) that a null hypothesis of the sort embodied in Equation 2 is false. Such a finding provides no information about the form of the underlying pattern of population means, which is presumably what is important for making scientific conclusions.

There are several ways of dealing with this problem. One way is to use post hoc tests, whereby decisions whether to reject or fail to reject are made for null hypotheses involving particular pairs of means. But post hoc tests have problems. First, within the hypothesis-testing framework, the more such tests are carried out, the greater is the probability of committing at least one Type I error. It is possible to adjust for this problem, but only at the expense of raising the probability of committing a Type II error. Second, and more generally, post hoc tests focus on specific pairs of means; when there are more than two conditions, these tests provide only an indirect way of assessing the entire pattern of means.

A second way to address the problem is via planned comparisons. Use of planned comparisons entails first generating a pattern of *weights*—one weight per experimental condition—that constitute the prediction of some experimental hypothesis about the overall pattern of population means. The correlation between the weights and the observed sample means then constitutes (essentially) a measure of how good the hypothesis is. Use of planned comparisons can be an informative and efficient process. The problem with planned comparisons, however, is that, in practice, they are rarely used. I return to the topic of planned comparisons in a later section.

Power

The third problem with NHST has to do with lack of attention to statistical power. NHST has come to revolve critically around the avoidance of Type I errors, mainly because the probability of a Type I error (α) can be computed. In contrast, the probability of a Type II error (β), and concomitantly power ($1 - \beta$), usually cannot be computed because computation of β and power requires a specific, quantitative hypothesis (e.g., $\mu_2 = \mu_1 + 10$ in a two-condition experiment), and such quantitative hypotheses are exceedingly rare in the social sciences.

Lack of power analysis is particularly troublesome when an investigator concludes that some null hypothesis is true (rather than making the logically correct decision of failing to reject the null hypothesis).[7] In such a case, one cannot be sure whether there are in fact relatively small differences among the population means (a conclusion that is justifiable only if there is high power) or whether there may be large differences among population means that are undetected (a conclusion that is implied by low power). In short, with high power, an investigator may be justified in accepting the null hypothesis "for all intents and purposes," but the lower the power, the less acceptable is such a conclusion.

As noted, lack of power analysis often stems from the lack of quantifiable alternative hypotheses that characterizes the social sciences in general, and psychology in particular. Nonetheless, there are ways of conveying the overall state of statistical power in some experiment (particularly through use of confidence intervals, as is illustrated in an example in the section on "Alternatives").

The Artificial "Effects/Non-Effects" Dichotomy

A related problem is not, strictly speaking, a problem in the logic of NHST. Rather, it is a problem that arises because investigators, like all humans, are averse to making decisions that are both complicated and weak, such as "we fail to conclude that the null hypothesis is false." Rather, people prefer simple, strong decisions, such as "the null hypothesis is true." This fact of human nature fosters an artificial dichotomy that revolves around the arbitrary nature of the .05 α level.

Most people, if pressed, will agree that there is no essential dif-

ference between, say, finding that $p = .050$ and finding that $p = .051$. However, investigators, journal editors, reviewers, and scientific consumers often forget this and behave as if the .05 cutoff were somehow real rather than arbitrary. Accordingly, the world of perceived psychological reality tends to become divided into "real effects" ($p \leq .05$) and "non-effects" ($p > .05$). Statistical conclusions about such real effects and non-effects made in Results sections then somehow are sanctified and transmuted into conclusions that endure into Discussion sections and beyond, where they insidiously settle in and become part of our discipline's general knowledge structure. The mischief thereby stirred up is incalculable. For instance, when one experiment shows a significant effect ($p \leq .05$), and an attempted replication shows no significant effect ($p > .05$), a "failure to replicate" is proclaimed. Feverish activity ensues, as Method sections are scoured and new experiments run, in an effort to understand the circumstances under which the effect does or does not show up—and all because of an arbitrary cutoff at the .05 α level. No wonder there is an epidemic of "conflicting" results in psychological research! This state of affairs is analogous to a chaotic phenomenon in which small initial differences lead to enormous differences in the eventual outcomes. In the case of data analysis, chaos is inimical to understanding, and it is more appropriate that similar results (e.g., $p = .050$ and $p = .051$) yield similar conclusions than that similar results yield entirely different conclusions.

The Hypothesis-Testing Tail Wags the Theory-Construction Dog

Central to NHST is computation of p, the probability of the data given the null hypothesis. However, this probability is difficult or impossible to compute unless routine, simplifying assumptions are made about the nature of the psychological processes under consideration. These assumptions then insinuate themselves into, and become integral to, much of psychological theory. Some of the most common such assumptions are these:

1. The dependent variable is obtained by adding up numerical "effects" (i.e., is a linear combination of those effects)—effects due to the independent variables, to the interactions among the independent variables, and to various sources of "error."
2. The errors are distributed according to Gaussian (normal) distributions.
3. The variances of these error distributions are equal across various conditions.

Thus, the nature of the data-analysis technique generally dictates the nature of psychological theory, which, in turn, engenders strong biases against formulating theories incorporating other perhaps more realistic or interesting assumptions. Accordingly, psychological theory becomes generic linear-model theory, and a lot of potential for insight is lost. One might summarize the situation as "Off-the-shelf assumptions produce off-the-shelf conclusions."

For example, suppose an experiment is designed to investigate whether the rate of forgetting depends on degree of original learning. In this experiment, word lists are taught to subjects whose recall is tested following intervals that vary from 0 to 5 days. There are two groups of subjects. The high-learning subjects are allowed to learn the lists to a criterion level of 100% correct; the low-learning subjects learn the lists to a lower criterion level. The major data from this experiment are forgetting curves of the sort shown in Figure 1, which plots memory performance as a function of retention interval.

The default data-analysis procedure in such an experiment would be to carry out a two-way analysis of variance (ANOVA). Let us suppose that the experimental power is sufficiently great that the ANOVA reveals statistically significant main effects of both learning level and retention interval, along with a significant interaction. It is evident in Figure 1 that the interaction reflects a shallower "slope" for the low-learning than for the high-learning condition (i.e., the change in memory performance over the 5 days is greater for the high-learning group, so that the vertical difference between the curves becomes smaller with increasing retention interval). The typical conclusion issuing from these observations would be that forgetting is slower following low learning than following high learning (this logic was used by Slamecka & McElree, 1983[8]).

This standard data-analysis procedure, along with the concomitant conclusion, would mask a very interesting regularity in the data, however: As is indicated by the horizontal lines on the figure, the horizontal difference between the high-learning and low-learning forgetting curves is constant. As I have shown (Loftus 1985; see also Loftus & Bamber, 1990), such horizontal equality is, under very general assumptions, a necessary and sufficient condition to infer equal high- and low-learning forgetting rates. Indeed, the Figure 1 curves were generated from exponential decay equations of the following form:

$$\text{Low learning: } p = e^{-0.3t}$$

$$\text{High learning: } p = e^{-0.3(t+2)}$$

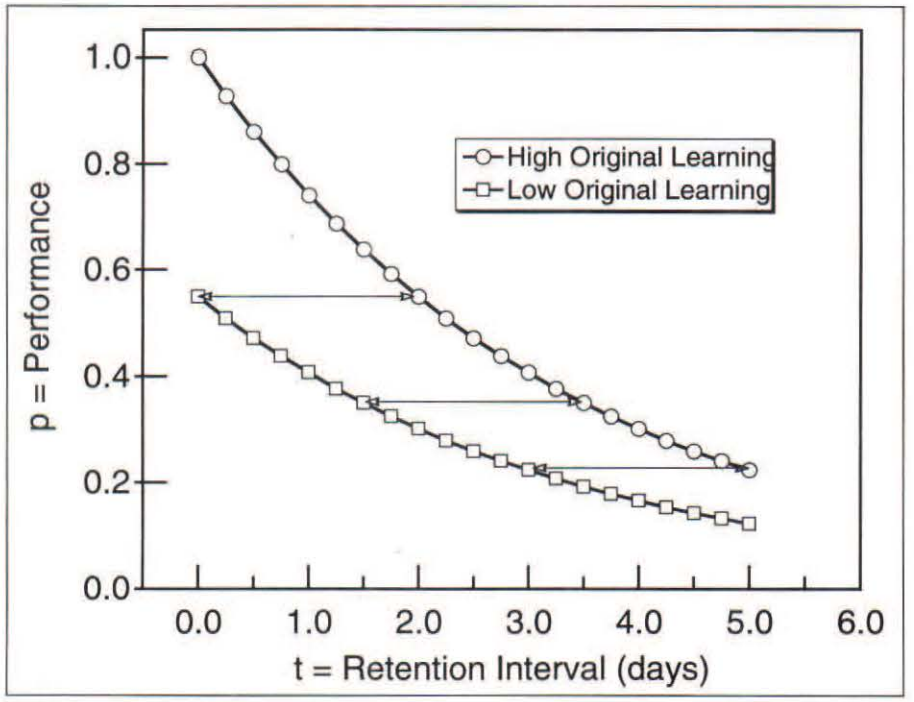

Fig. 1. Forgetting curves following high and low learning. Horizontal lines are meant to indicate that the two curves are horizontally parallel.

where p is performance and t is forgetting time (in days). The equal forgetting rates are expressed by the same exponential decay parameter (0.3) in both equations.

In this example, the principal finding is (or should be) that the horizontal difference between the two curves is constant. It is this finding that implies no difference between the forgetting rates of the high- and low-learning groups. However, investigators working within the hypothesis-testing framework would tend to miss this critical regularity, or would dismiss it, for at least two reasons. First, it is not immediately obvious how a standard significance test that is relevant to the finding could be carried out,[9] and without a significance test, a finding is not generally deemed "valid." Second, the logic of the linear model within which standard ANOVA is couched focuses on differences between the dependent variable at a fixed level of the independent variable (vertical differences), rather than differences between the independent variable at a fixed level of the dependent variable (horizontal differences).

NHST Provides Only Imprecise Information About the Validity of the Null Hypothesis

The final problem, which has been hammered at by Bayesian statisticians (e.g., Berger & Berry, 1988; see also Cohen, 1994) for decades, is this: By convention, one rejects some null hypothesis when

(1) p(observed data|null hypothesis) $< .05$,

but rejecting the null hypothesis implies the conclusion that

(2) p(null hypothesis|observed data) is small.

(What else could be meant by the phrase "reject the null hypothesis"?) But without additional information, there is no logical basis for concluding the validity of Statement 2 given the finding embodied in Statement 1. Indeed, the probability of the null hypothesis given the data could be shown to be anything given suitable assumptions about the prior (predata) probability that the null hypothesis is true. Without specific assumptions about this prior probability, the exact probability of the null hypothesis given the observed data is unknown. In short, the common belief that the precise quantity .05 refers to anything meaningful or interesting is illusory.

ALTERNATIVES

I now suggest four (by no means mutually exclusive) alternatives to traditional NHST. My major goal in making each of these suggestions is simple and modest: to increase our ability to understand what a data set is trying to tell us. These techniques are not fancy or esoteric. They are just sensible.

Plot Data Rather Than Presenting Them as Tables-Plus-*F*-and-*p*-Values

In a previous article (Loftus, 1993b), I described some fictional data collected by a fictional psychologist named Jennifer Loeb. The story went as follows. Loeb was interested in memory for visual material and carried out a task in which visual stimuli were displayed and then recalled. There were three independent variables in Loeb's experiment, all varied at the time of stimulus presentation: stimulus exposure duration (eight values, ranging from 20 ms to 230 ms), verbal encoding of the stimuli (whether naming them was prohibited or required), and uncertainty about where the stimuli would appear (high or low). Based

on a specific theory, Loeb had three predictions. First, she predicted her performance measure to be an increasing linear function of stimulus exposure duration. Second, she predicted the slope of this function to be higher with than without verbal encoding. Third, she predicted the slope to be higher with low than with high spatial uncertainty.

Table 1, which shows a common way of presenting data like Loeb's, lists the values of her performance measure for all 32 (8 × 2 × 2) conditions. Accompanying NHST results—long compendia of F ratios and p values corresponding to the main ANOVA plus subsidiary tests—are generally provided as part of the text (often spanning many tedious pages).

Most people find such tabular-cum-text data presentation difficult to assimilate. That is not surprising. Decades of cognitive research, plus millennia of common sense, teach us that the human mind is not designed to integrate information that is presented in this form. There is too much of it, and it cannot be processed in parallel.

An alternative way of presenting Loeb's data is shown in Figure 2. Performance is plotted as a function of exposure duration for the verbal-encoding conditions (top panel) and the no-verbal-encoding conditions (bottom panel). The two curves within each panel are for the low-uncertainty and high-uncertainty conditions. Best-fitting linear functions are drawn through the data points. With the data presented like this, one can acquire in a glance—or at most, a couple of glances—the same information that it would have taken practically forever to get out of Table 1. A picture really *is* worth a thousand words.[10]

Provide Confidence Intervals

The Figure 2 plot indicates that Loeb's obtained pattern of sample means confirms her predictions pretty well. The curves are generally linear, verbal encoding yields a higher slope than no verbal encoding, and low uncertainty yields a higher slope than high uncertainty.

However, this plot provides no indication of the sort of error variance that is typically included as part of an ANOVA. Figure 3 remedies this deficiency: It shows the same data along with 95% confidence intervals (i.e., intervals that show the range of values within which the true population means lie with 95% probability). In this within-subjects design, the confidence intervals were computed using a method Masson and I have described elsewhere (Loftus & Masson, 1994).

Figure 3 illustrates my second suggestion, which is to put confidence intervals around all sample statistics that are important for making conclusions. Figure 3 provides most of the crucial information about Loeb's data. The observed pattern of sample means provides the best estimate of the underlying pattern of population means upon which conclusions should be based. The provision of confidence intervals allows the reader to assess the degree of statistical power: The smaller the confidence intervals, the greater the power. Power, in contrast to its profoundly convoluted interpretation within the hypothesis-testing framework, can be simply interpreted in this graphic presentation as an indication of how seriously the observed pattern of sample means should be taken as a reflection of the underlying pattern of population means.

One could analyze these data further by, for instance, computing the slope for each of the four Verbal Encoding × Uncertainty conditions for each subject. These four slopes could then be plotted along with *their* confidence intervals. One could go still further by, for instance, computing mean slope differences along with *their* confidence intervals. Such procedures correspond to graphic illustrations of various kinds of interactions. Creative use of such procedures allows one to jettison NHST entirely.

Confidence Intervals as a Guide to Accepting the Null Hypothesis

The provision of confidence intervals is particularly useful when one wants to accept some null hypothesis "for all intents and purposes." Another fictional data set (also introduced in Loftus, 1993b) involves a clinician whom I call Christopher Sanders. In this account, Sanders developed a clinical technique, designed to decrease agoraphobia, that is cheaper than the generally used standard technique. Sanders ran a

Table 1. *Performance as a function of exposure duration for four conditions*

	Exposure duration (in milliseconds)							
Condition	20	50	80	110	140	170	200	230
NVE,HU	0.287	0.503	0.843	1.005	1.468	1.664	2.102	2.257
VE,HU	0.461	1.192	1.399	2.360	3.008	3.236	3.908	4.649
NVE,LU	0.099	0.536	1.192	1.461	1.626	2.048	2.657	2.874
VE,LU	0.683	1.475	2.822	3.747	4.863	5.397	6.861	7.849

Note. NVE = no verbal encoding; VE = verbal encoding; HU = high spatial uncertainty; LU = low spatial uncertainty.

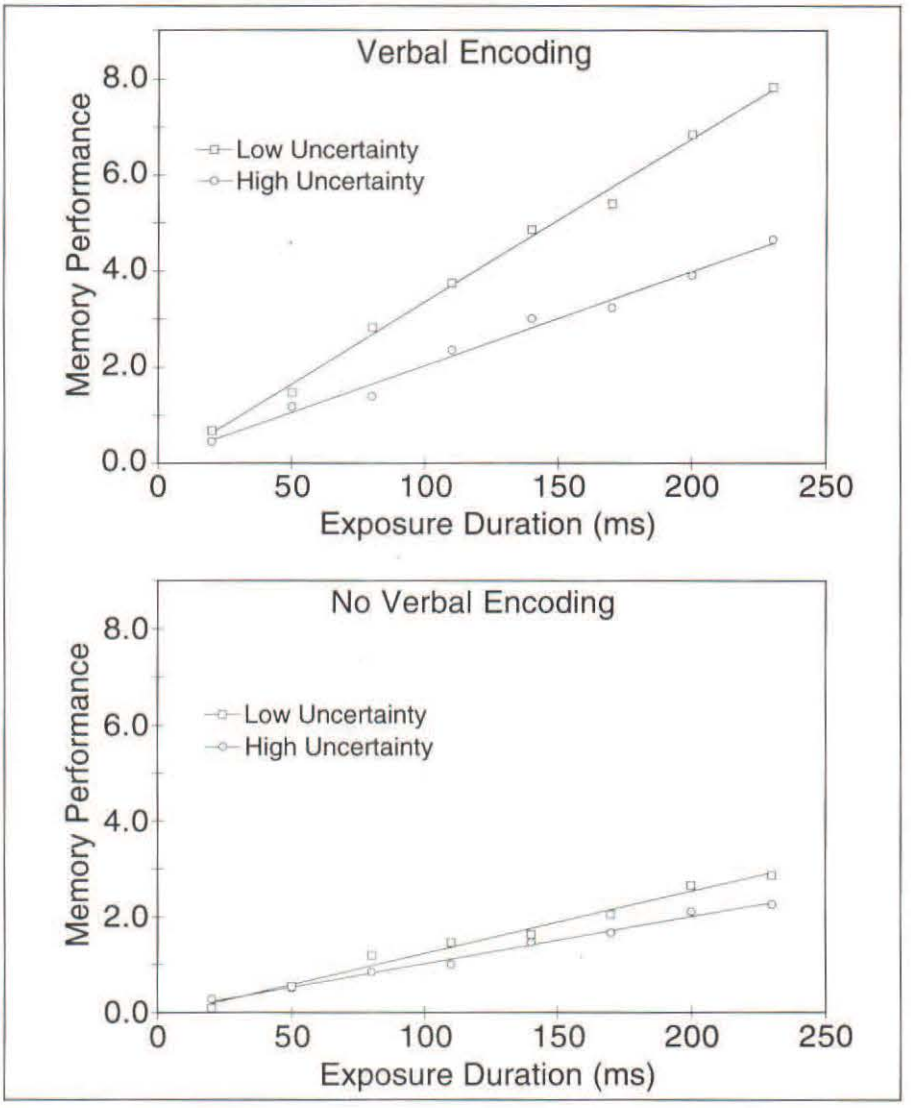

Fig. 2. Loeb's data.

simple experiment to compare his technique with the standard technique. In this experiment, 40 agoraphobic individuals were randomly assigned to be administered either the standard or the Sanders treatment. A year later, each individual's agoraphobia was assessed on a 10-point scale. Sanders's hope was that there would be no difference between the two treatments, in which case his treatment, being cheaper, would presumably be preferred to the standard treatment.

Sanders got his hoped-for result, and reported it thusly: "The mean agoraphobia scores of the standard and the Sanders groups were 5.05 and 5.03. The difference between the two groups was not statistically significant, $p > .05$." Sanders went on to conclude that the cheaper Sanders technique was therefore the preferred one.

What is implied by the phrase "not significantly different" in Sanders's report? We cannot tell, because Sanders provided no indication of statistical power, evaluation of which would be critical for justifying his accepting the null hypothesis of no treatment difference.

Sanders's report is consistent with many possible outcomes, two of which are presented in Figure 4. In the top panel, small confidence intervals reflect high experimental power. If this were Sanders's outcome, acceptance of the null hypothesis would be reasonable: It is readily apparent that the actual difference between the Sanders and standard population means must be quite small. In contrast, in the bottom panel, large confidence intervals reflect low statistical power. If this were Sanders's outcome, acceptance of the null hypothesis would be unconvincing: The actual difference between the population means for the two treatments could vary widely.

There is a noteworthy epilogue to this story: When pressed at a professional conference, Sanders further defended his acceptance of the null hypothesis by pointing out that the t value he obtained in his t test was very small—"Practically zero!" he declared proudly. And so it was that Sanders committed the common error of equating smallness of the test statistic with permissibility of accepting the null hypothesis. Ironically, it can be shown easily that, given a particular mean difference, the smaller the t value, the lower is power—and hence, the less appropriate it is to accept the null hypothesis.

Plus Ça Change, Plus C'est la Même Chose

The notion of using confidence intervals in this way—as a guide to accepting a null hypothesis "for all intents and purposes"—is not new. Thirty-four years ago, the following suggestion appeared in the pages of the *Psychological Review*.

> In view of our long-term strategy of improving our theories, our statistical tactics can be greatly improved by shifting emphasis away from overall NHST in the direction of statistical estimation. For example [when testing a presumed null pre-treatment difference between two groups, an investi-

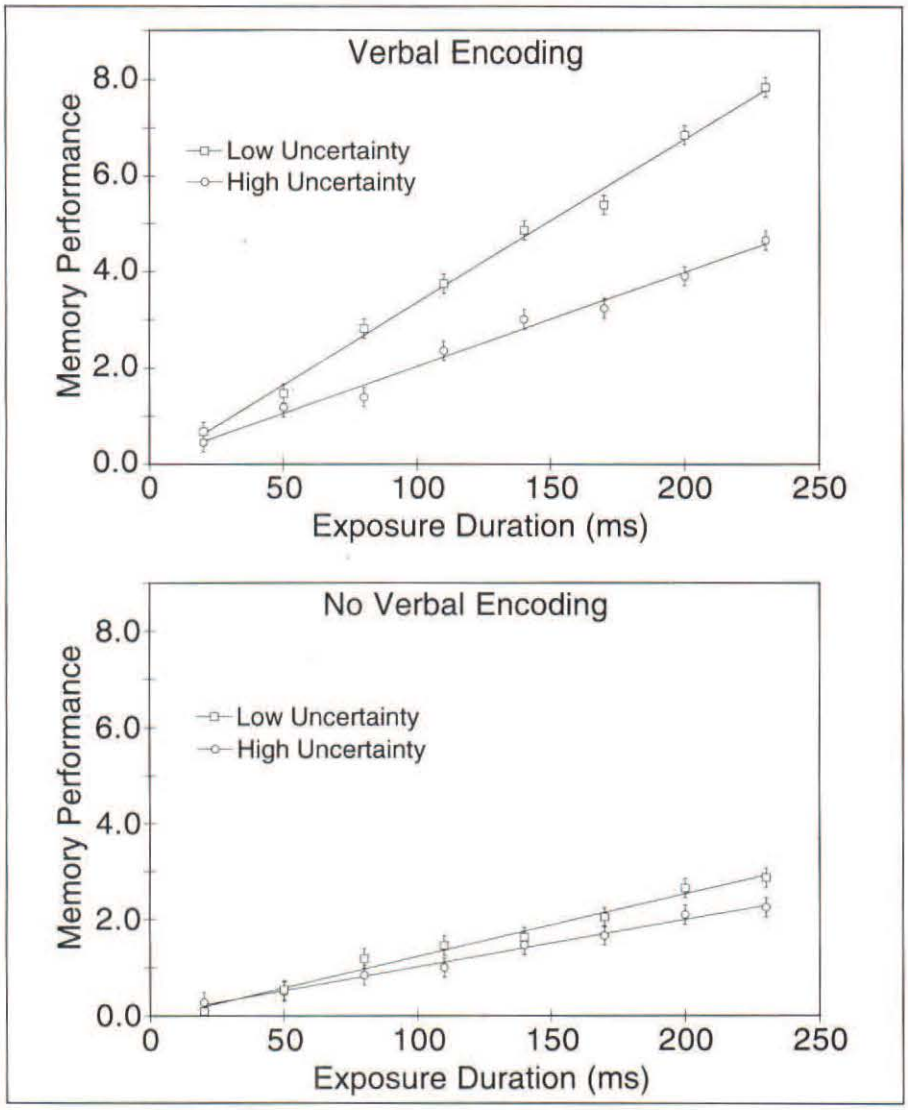

Fig. 3. Loeb's data plotted with confidence intervals.

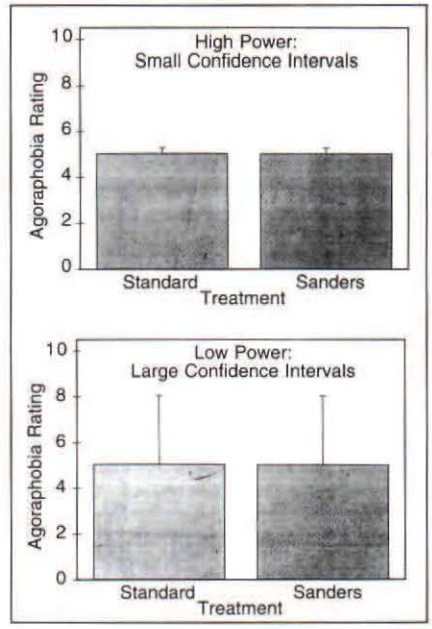

Fig. 4. Two possible outcomes of Sanders's experiment, illustrating high and low power.

gator] would do better to obtain a . . . confidence interval for the pretreatment difference. If the interval is small and includes zero, [the investigator] is on fairly safe ground; but if the interval is large, even though it includes zero, it is immediately apparent that the situation is more serious. In both cases, H_0 would have been accepted. (Grant, 1962, p. 57)

Confidence Intervals and Statistical Significance

A section on confidence intervals would be incomplete without a discussion of the question: How do you make a decision about whether some variable has an effect by just looking at confidence intervals?[11] The answer is, you cannot.[12] This, I assert, is an advantage rather than a disadvantage of using plots plus confidence intervals rather than depending on NHST. As I have already argued, a major difficulty with NHST is that it reduces data sets into a series of effect/no-effect decisions, and this process, artificial as it is, leads the field astray in many ways. It imposes the illusion of certainty on a domain that is inherently ambiguous. Simply showing data, with confidence intervals, provides a superset of the quantitative information that is provided by a hypothesis-testing procedure, but it does not foster the false security embodied in a concrete decision that is based on a foundation of sand.

Meta-Analysis and Effect Size

An increasingly popular technique is that of meta-analysis (e.g., Schmidt, 1996; Rosenthal, 1995). Meta-analysis entails considering a large number of independent studies of some phenomenon (e.g., gender differences in spatial ability) and (essentially) averaging the observed effects across studies to arrive at an overall effect. This technique is particularly useful when two conditions are being compared, but trickier when the question under investigation involves more than two conditions.

As an illustration, consider the question of gender differences in spatial ability. Suppose that, in fact, there is some difference in spatial ability between the population of males and the population of

females. The direction and magnitude of this difference might be investigated in many (say, 25) separate studies. Each individual study produces some observed gender difference (presumably the mean male spatial score minus the mean female spatial score) that constitutes that experiment's estimate of the mean difference between the populations. Meta-analysis would entail averaging all 25 reported differences, thereby arriving at a single estimated difference that is (roughly speaking) five times as accurate as any of the individual estimates.

One problem with this technique is the number that should be extracted from each of the individual studies. One could simply use the raw difference, whatever it might be. The problem with this approach is that different studies presumably used somewhat different measures of spatial ability, different ways of carrying out the experiment, different populations of subjects, and so on; thus, the raw measures would not be comparable across the studies. The typical solution is to compute *effect size*, which, in its simplest incarnation, is the mean difference observed in a given experiment divided by the obtained estimate from that experiment of the population standard deviation. It is these effect sizes that are then averaged to arrive at an overall estimate of the population effect size.

Planned Comparisons (Contrasts)

My final suggestion, which I alluded to earlier, is to carry out planned comparisons on a data set. Planned comparisons is a technique that has been strongly advocated and clearly described by many investigators (see, e.g., Abelson, 1995). In my opinion, however, it is a technique that is surprisingly underutilized.

Briefly, carrying out a planned comparison involves the following steps.

1. First, one generates a quantitative hypothesis (even a relatively simple one will do) about the underlying pattern of population means corresponding to the conditions in some experiment. Suppose, for example, that a researcher is investigating the relation between problem-solving time and alcohol consumption. The researcher designs an experiment in which subjects are assigned to one of five groups that differ in amount of alcohol consumed—0, 1, 2, 3, or 4 oz—and problem-solving time is measured for each group. A hypothesis to be tested is that problem-solving time increases linearly with number of ounces of consumed alcohol.

2. Next, hypothesis in hand, the researcher generates weights—one weight per condition—that correspond to the hypothesized pattern of means. One constraint is that the weights must sum to zero; thus, in this example, appropriate weights representing the linearity hypothesis would be −2, −1, 0, 1, and 2.

3. The experiment is carried out, and the means—in this example, the five mean problem-solving times—are computed.

4. Finally, the researcher (essentially) computes an over-conditions correlation between the weights and the sample means. The magnitude of the Pearson r^2 that emerges reflects the goodness of the hypothesis. Various other more sophisticated procedures can also be carried out, but the nature of these procedures is beyond the scope of this article.

CONCLUSIONS

I have tried to provide a variety of reasons why NHST, as typically utilized, is barren as a means of transiting from data to conclusions. I have tried to provide some examples of techniques—standard techniques, not bizarre or fancy ones—to replace standard NHST. These techniques are as follows: first, to plot the data rather than putting them in tabular form; second, to put confidence intervals around important sample statistics; third, to use meta-analysis; and fourth, to use planned comparisons. These techniques are all designed to assist in the ultimate goal of understanding what it is that some data set is trying to tell us.

This article is titled "Psychology Will Be a Much Better Science When We Change the Way We Analyze Data." I hope that my arguments make it clear why I believe this to be true. I believe that in order for any science to progress satisfactorily, its primary data-analysis techniques must provide genuine insight into whatever phenomena its practitioners set out to investigate. The primary data-analysis technique of psychology—NHST—does not, as I have tried to demonstrate, meet this criterion.

Acquisition of insight is often difficult in the social sciences, which are cursed with large numbers of uncontrollable variables, and hence error variance that has to be dealt with somehow. I believe that, historically, social scientists have embraced NHST procedures because they provide the appearance of objectivity. These procedures may indeed be objective in the sense that they provide rules for making scientific decisions. But such objectivity is not, alas, sufficient for insight. I believe that these rules provide only the

illusion of insight, which is worse than providing no insight at all.

Acknowledgments—The writing of this manuscript was supported by a grant from the National Institute of Mental Health to the author. I thank Nelson Cowan, Gerd Gigerenzer, David Irwin, Elizabeth Loftus, Michele Nathan, and especially Emanuel Donchin for their very insightful and helpful comments on early versions of this manuscript.

Notes

1. Address correspondence to Geoffrey Loftus by e-mail: gloftus@u.washington.edu.

2. Actually, NHST can be applied to any population parameter. I use means here because means are by far the parameter of greatest concern in social science experiments.

3. A brief note of clarification is in order here. The entity I have referred to as p, which is computed from the data, is Fisher's exact level of significance (Fisher, 1925). The related entity known as α is the Neyman-Pearson probability of a Type I error, which is decided upon before the data are collected. The Neyman-Pearson approach to statistics has been "hybridized" with what is known as the Fisher approach over the years (as has been splendidly described by Gigerenzer et al., 1989, pp. 78–109), even though these two approaches are quite different. It is the resulting mishmash that has been almost universally taught as "the statistical method" over the past half century. A detailed analysis of this issue is beyond the scope of this article, but Gigerenzer et al. made a convincing case that the confusion of the Fisher and Neyman-Pearson approaches is responsible for much of what has gone astray with modern statistical practice. The description of NHST that I provide here—and that I inveigh against—is, essentially, a description of this commonly used hybridized approach.

4. A sample of these writings is, in chronological order: Tyler (1935), Jones (1955), Nunnally (1960), Rozeboom (1960), Grant (1962), Bakan (1966), Meehl (1967), Lykken (1968), Carver (1978), Meehl (1978), Berger and Berry (1988), Gigerenzer et al. (1989), Rosnow and Rosenthal (1989), Cohen (1990), Meehl (1990), Loftus (1991), Carver (1993), Cohen (1994), Loftus and Masson (1994), Maltz (1994), and Schmidt (1996).

5. This situation may finally be changing. Symposia at both the 1996 American Psychological Society (APS) convention and the 1996 American Psychological Association (APA) convention have aired the shortcomings of NHST as the primary data-analysis technique in the social sciences. (The papers from the APS symposium will appear in a Special Section of the January 1997 issue of *Psychological Science*, Vol. 8.) An APA task force has been set up to study the value of NHST, and at least one journal editor has tried to discourage NHST (Loftus, 1993a, which provoked the observation by Greenwald, Gonzalez, Harris, & Guthrie, 1996, that every recent empirical article in Loftus's journal, *Memory & Cognition*, has used NHST nonetheless).

6. One caveat is in order here. There are some experiments in which a null hypothesis could genuinely be true. A good example (attributable to Greenwald et al., 1996) is a qualitative null hypothesis such as that a defendant in a murder case is actually the murderer. In such a case, the null hypothesis could certainly be true, and rejecting it (say, based on DNA matches) would be a meaningful conclusion. However, these kinds of experiments are the exception rather than the rule in the social sciences.

7. As indicated earlier, it is usually known on a priori grounds that a null hypothesis cannot be literally correct. However, as I discuss in more detail in a later section, given sufficient power along with close-to-equal sample means, one can justifiably accept a null hypothesis "for all intents and purposes."

8. Actually, Slamecka and McElree found no significant interaction between degree of original learning and retention interval, and hence concluded that forgetting rate did not depend on degree of original learning. However, different data sets using the same general paradigm show different forms of interactions that would lead variously to the conclusion that forgetting is faster for high learning, that forgetting is slower for high learning, and that rates of forgetting for high and low learning do not differ (a meta-finding that should, in and of itself, provoke suspicion that something is fundamentally amiss). In any event, it is the logic of the data-analysis technique, not the conclusion, that is primarily at issue here.

9. This is not to say such a test could not be invented; it simply is not part of general statistical knowledge or (what is probably more important) part of present statistics computer packages.

10. One might argue that tables are useful when a reader needs exact values of the data points. However, such situations are quite rare, and when they do occur, exact values are obtainable from the investigators—a process that is particularly easy in these days of electronic communication.

11. In numerous statistics classes and in other forums in which I have discussed this issue, a question that I can absolutely count on is, In a two-condition situation, what is the relation between error-bar overlap and the reject/fail-to-reject decision?

12. At least, not usually. In a two-group design, finding nonoverlapping 95% confidence intervals implies that a two-tailed, $\alpha = .05$ t test would lead to a conclusion of "statistically significant." However, if the error bars do overlap, or if there are more than two conditions in the experiment, the relation between the pattern of confidence intervals and statistical significance is not immediately apparent.

References

Abelson, R.P. (1995). *Statistics as principled argument*. Mahwah, NJ: Erlbaum.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437.

Berger, J.O., & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*, 159–165.

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, revisited. *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.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.

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., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect sizes and p-values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.

Hays, W. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt.

Jones, L.V. (1955). Statistics and research design. *Annual Review of Psychology, 6,* 405–430.

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

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

Loftus, G.R. (1993a). Editorial comment. *Memory & Cognition, 21,* 1–3.

Loftus, G.R. (1993b). Visual data representation and hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers, 25,* 250–256.

Loftus, G.R. (1995). Data analysis as insight. *Behavior Research Methods, Instruments, & Computers, 27,* 57–59.

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

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

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). Theoretical 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, 66*(Suppl. 1), 195–244.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20,* 641–650.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

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

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

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

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.

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

The Water-Level Task: An Intriguing Puzzle

Ross Vasta and Lynn S. Liben[1]

Department of Psychology, SUNY Brockport, Brockport, New York (R.V.), and Department of Psychology, The Pennsylvania State University, University Park, Pennsylvania (L.S.L.)

Take a moment to look at Figure 1. It presents one of several variations of an intriguing problem known as the water-level task (WLT). The correct response to the problem is to draw a horizontal line across the bottle, reflecting the general principle that the surface of a liquid is invariantly horizontal regardless of the orientation of its container. Variations of the task have included presenting the tilted bottle alone, using real containers rather than drawings, and asking subjects whether a waterline in a tilted container looks "correct" (rather than having them draw a line).[2]

The WLT might appear to be a simple problem. In reality, researchers have found that a surprisingly large proportion of adolescents and adults draw slanting lines in the tilted bottles (often with considerable confidence!), and are unable to articulate or identify the physical principle underlying the task. Determining which subjects are most likely to make errors, and why they do so, has been a 30-year scientific puzzle that continues to challenge investigators.

In this article, we begin by tracing the WLT to its source and original purpose—Piaget's work on children's spatial development. We then examine how the task provided an inadvertent battleground for the theoretical debate surrounding gender differences that emerged during the 1970s. Finally, we consider current attempts to explain the fascinating data that continue to be generated by the WLT, and we suggest some directions future research might take.

ORIGINS OF THE PROBLEM

The WLT was developed by Piaget and Inhelder (1948/1956) as part of their investigation of children's emerging spatial concepts. Piaget and Inhelder proposed that

Recommended Reading

Kalichman, S.C. (1988). Individual differences in water-level performance: A component skills analysis. *Developmental Review, 8,* 273–295.

Liben, L.S. (1991). The Piagetian water-level task: Looking beneath the surface. In R. Vasta (Ed.), *Annals of child development: Vol. 8* (pp. 81–144). London: Kingsley.

Pascual-Leone, J., & Morra, S. (1991). Horizontality of water level: A neo-Piagetian developmental review. In H.W. Reese (Ed.), *Advances in child development and behavior: Vol. 23* (pp. 231–276). San Diego: Academic Press.