Comment

Contents

The Inscrutable Null Hypothesis

Warren W. Tryon
Fordham University

Hagen's (January 1997) article praising the null hypothesis statistical test (NHST) also cited literature critical of it and recognized that NHST "has been misinterpreted and misused for decades" (p. 22). NHST criticism goes back farther than the 30 years acknowledged by Hagen. Pearce (1992) reported that criticism of NHST began immediately with Fisher's introduction of it in 1925. Despite continuous critical commentary over the past 72 years, NHST became the primary method of data analysis in the social sciences.

A principal human factors requirement of any viable data analytic procedure, regardless of its other merits or demerits, is that it can be correctly calculated and interpreted. Widespread access to commercial statistical packages indicates that NHST calculations reported in the literature are probably correct. However, substantial reasons seriously question whether NHST results have been, are, or can be correctly interpreted consistently by most investigators.

Carver (1978) identified several misinterpretations of NHST results and reported that practices were unchanged 15 years later (Carver, 1993). Dar, Serlin, and Omer (1994) surveyed three decades of NHST misuse published in the *Journal of Consulting and Clinical Psychology* between 1967 and 1988. Cohen (1994) cited texts written by six prominent psychometricians that misinterpret NHST results. McMan (1995) found substantial NHST errors in most of 24 introductory psychology textbooks published between 1965 and 1994. Hagen's (1997) need to reprove three of Cohen's (1994) NHST criticisms indicates that even a prominent author of multiple statistics texts seemingly cannot "correctly" interpret NHST results. How much more susceptible to misinterpretation are the vast majority of other less well quantitatively trained psychologists?

Regardless of the technical merits or demerits of NHST, the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are undoubtedly substantial and may be the strongest reason for adopting other data analytic methods. Hagen's (1997) praise of NHST may be supportable on purely technical grounds but is unfortunate if it prolongs primary reliance on NHST to evaluate quantitative difference and equivalence given the prominent human factors problem of widespread and intractable interpretation errors. Alternative methods are available for these purposes that are far less subject to misinterpretation. The science of psychology can only benefit by supplementing, if not replacing, NHST practices with these methods.

REFERENCES

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

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75–82.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

McMan, J. C. (1995, August). *Statistical significance testing fantasies in introductory psychology textbooks.* Paper presented at the 103rd Annual Convention of the American Psychological Association, New York.

Pearce, S. C. (1992). Introduction to Fisher (1925): Statistical methods for research workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics, Vol. 2: Methodology and distributions* (pp. 59–65). New York: Springer-Verlag.

Correspondence concerning this comment should be addressed to Warren W. Tryon, Department of Psychology, Fordham University, Bronx, NY 10458-5198. Electronic mail may be sent to wtryon@murray.fordham.edu.

Significance Testing: Is There Something Better?

Robert E. McGrath
Fairleigh Dickinson University

In the article "In Praise of the Null Hypothesis Statistical Test" (NHST), Hagen (January 1997) did an admirable job of reminding readers that NHST represents a brilliant and useful innovation, with relevance for research settings far more sophisticated than those originally considered by its creator. However, it is important to note that even this very supportive article does not offer a strong case for its continued use as the primary inferential strategy in psychology. I address five particular aspects of the article.

First, Hagen (1997) suggested that "if we are content to equate the $P(H_0)$ with a subjective degree of belief, or level of confidence, then the NHST does, indeed, tell us what we want to know" (p. 19). Instead, what Hagen demonstrated is what a Bayesian analysis of NHST results can reveal about

0003-066X/98/$2.00
Vol. 53, No. 7, 796–807

$P(H_0)$. This is not a trivial distinction. Bayes's theorem is rarely taught to psychologists or used as an adjunct to NHST. This is in part the fault of Fisher (1937) himself, who specifically opposed the use of Bayes's theorem in this context. It is worth noting that the same argument raised by Hagen in support of NHST was originally introduced by critics of the method who took Fisher at his word about how NHST was supposed to proceed (see Oakes, 1986).

Second, Hagen (1997) responded to the popular belief that an effect size exactly equal to zero is unlikely, rendering an analysis aimed at evaluating whether the effect equals zero absurd. His argument seems to be that although in any one study the effect will never equal zero, there is no reason to believe these discrepancies will not even out across studies, leaving the null hypothesis true at the level of the population.

Although it is true as noted that a zero effect is not impossible, it is highly unlikely that an effect will exactly equal zero in anything less than the most well-controlled studies. As noted by Cohen (1994), there is no reason to expect that population correlations between uncontrolled variables exactly equal zero, so the use of no-effect significance tests in any observational study is suspect. Even in well-controlled true experiments, there are often nonrandom nuisance variables inherent to the experimental design that cannot be perfectly controlled (Campbell & Stanley, 1963). Hagen (1997) himself stated that "Tukey's (1991) comment that the effects of A and B are always different can stand. But it does not necessarily follow that the null hypothesis will always be vulnerable to those effects" (p. 21). This is a far cry from saying that the null hypothesis can usually be considered invulnerable to these effects, a statement that would be more consistent with recommending the widespread use of NHST.

Third, NHST has been criticized because as a system for the testing of propositions, it does not demonstrate the same level of logical validity as the modus tollens. Hagen (1997) accepted this critique but responded by suggesting that evaluations of scientific propositions rarely demonstrate the highest level of logical validity. It is an interesting argument but again begs the question of whether there are more logically justifiable methodologies.

Fourth, Hagen (1997) responded to Cohen's (1994) and Schmidt's (1996) preference for confidence intervals over NHST by suggesting that confidence intervals are no better than NHST for the purpose of testing null hypotheses. This is true, but it ignores the primary reason for preferring confidence intervals. Under NHST, the basic question in primary research is "Based on this sample, what is our best guess about whether or not ρ equals 0?" The computation of confidence intervals allows for a much more interesting question: "Based on this sample, what is our best guess about the value of ρ?" This represents a fundamental change in the way that the analytic process is conceptualized. It is only if one fails to look beyond the limits of NHST that confidence intervals and NHST appear to be equivalent strategies.

Finally, Cohen (1994) clearly did not intend his article to be a comprehensive review of the problems associated with significance testing. By narrowly focusing on the arguments raised by Cohen, Hagen (1997) ignored the bulk of the criticisms leveled against the method. These criticisms include, among others, the logical problems associated with making a binary decision, the inevitably arbitrary element in the selection of alpha, the negligence of sample size issues fostered by Fisher's (1937) model of NHST, and obstacles to the accumulation of knowledge in psychology created by the use of NHST (Schmidt, 1996). Hagen's conclusion that "I have tried to point out . . . that the logic underlying statistical significance testing has not yet been successfully challenged" (p. 22) seems particularly excessive given the limited range of his response.

Hagen (1997) as well as Frick (1996) offered good, albeit incomplete, responses to those who would suggest NHST is useless or hopelessly, logically flawed. However, I do not think the question has ever really been "Is it useless?" but rather "Is there something better?" This question deserves much closer scrutiny than is possible here, but a popular opinion holds that interval estimation represents a superior strategy to NHST in many ways. Given all that has been gained through its use, I think it is very appropriate to praise the brilliance of NHST, but having done so, perhaps it is time to bury it.

REFERENCES

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

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

Fisher, R. A. (1937). *The design of experiments.* London: Oliver & Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester, England: Wiley.

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

Correspondence concerning this comment should be addressed to Robert E. McGrath, School of Psychology, T110A, Fairleigh Dickinson University, Teaneck, NJ 07666. Electronic mail may be sent to mcgrath@alpha.fdu.edu.

In Praise of Value Judgments in Null Hypothesis Testing . . . and of "Accepting" the Null Hypothesis

Robert G. Malgady
New York University

As Hagen (January 1997) acknowledged and as Cohen (1994) did before him, there has been considerable discourse on the merits and limitations of null hypothesis testing, dating back to Ronald Fisher (1935) himself. Nonetheless, as insightful as even Hagen's illumination of null hypothesis testing is, I believe two related issues have been obscured, if not neglected.

As most statisticians and philosophers of logic would say, one can reject a null hypothesis but can never accept or validate it by using the classical Fisherian (Fisher, 1935) procedure. I have argued elsewhere that in clinical research, this is fundamentally like burying one's head in the sand (Malgady, 1996). If the null hypothesis is not rejected in a statistical test, one certainly cannot assert that it has scientific validity, but behavior concerning the null hypothesis validates it because people act as if it were true. For instance, if a psychopharmacological researcher tests a new drug for treating major depression disorder, a null hypothesis might be that mean reduction of depressive symptomatology does not differ between an experimental (drug) condition and a placebo control. If this null hypothesis is not rejected, the researcher cannot lay scientific claim to its validity. But the obvious consequence of this decision is that, rightly or wrongly so, the drug will not be prescribed for persons with major depression disorder. Science dictates a conservative or skeptical stance—scientists don't believe in something until there is evidence of its truth within, of course, a comfortable margin of risk (e.g., probability of being in error < .05). Thus, there is a family of status quo null hypotheses composing a

belief system—essentially all the phenomena not believed to be true. Although these negative beliefs may not be scientifically valid or may never be subjected to scientific testing, they nonetheless govern behavior until someone rejects a null hypothesis here and there. For instance, I have a null hypothesis that peanut butter does not cure the common cold, although I cannot scientifically prove it. Although it is not scientifically valid, I do not give my children peanut butter when they catch a cold. This logic extends to professional practice and public policy in psychology. Rightly or wrongly, scientists "accept" null hypotheses when they behave as if they were true.

The second issue concerns which of two competing and mutually exclusive hypotheses is formulated as the null hypothesis. By convention (Hays, 1973), Type I errors (rejecting a true null hypothesis) are generally considered more serious than Type II errors (failing to reject a false null hypothesis). Scientists adhere rather rigidly to a maximum risk of Type I error equal to .05, and although .20 is recommended as a maximum risk of Type II error, Cohen (1994) estimated that, in practice, the prevalence of Type II errors is more like .50. Which type of error is more likely, therefore, depends entirely on which of the two competing hypotheses is premised as the hypothesis to be rejected. Hays argued that the more serious of the two types of errors should be determined first and then the corresponding hypothesis risking this error plays the null role "so that the abhorrent Type I error is very unlikely to be committed" (p. 368). I have argued elsewhere that such a decision preceding scientific hypothesis testing is predicated on what may be an unscientific and subjective value judgment (Malgady, 1996).

Although Hagen's (1996) article provides a lucid clarification of what can and cannot be learned from null hypothesis testing and points out some of the pitfalls and possible misinterpretations of Cohen's (1994) earlier article, both articles are constrained to the logic of deduction from an already formulated null hypothesis and the scientific conclusions that are validly drawn from one statistical decision or another. I believe that subjective value judgment preceding the construction of the null hypothesis is an obscure precursor of the scientific logic of null hypothesis testing and that the subsequent actions taken in professional practice as a result of failure to reject the null hypothesis constitute its acceptance. Both issues need more extensive consideration.

REFERENCES

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

Fisher, R. A. (1935). *The design of experiments.* New York: Hafner.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Hays, W. (1973). *Statistics for the social sciences.* New York: Holt, Rinehart & Winston.

Malgady, R. G. (1996). The question of cultural bias in assessment and diagnosis of ethnic minority clients: Let's reject the null hypothesis. *Professional Psychology: Research and Practice, 27,* 73–77.

Correspondence concerning this comment should be addressed to Robert G. Malgady, Program in Quantitative Studies, New York University, 200 East Building, 239 Greene Street, New York, NY 10003. Electronic mail may be sent to malgady@aol.com.

In Criticism of the Null Hypothesis Statistical Test

Ruma Falk
Hebrew University

The null hypothesis statistical test (NHST) should be praised (Hagen, January 1997) for its original intent, that is, for asking the question "Could this have been a coincidence?" When positing the null hypothesis (H_0) of no effect, we hope to reject it to eliminate the threat that our random sample's outcome might be a fluke of chance. Researchers and readers are universally plagued by this concern. Unfortunately, significance tests do not deliver the goods. To reject H_0, one needs to show that it had become unlikely by one's results. NHST fails to do so.

NHST is, in fact, a probabilistic imitation of modus tollens (or of the mathematical procedure of proof by contradiction). However, once the reasoning is made probabilistic, the inference is no longer valid (Cohen, 1994; Falk & Greenbaum, 1995). The following sentences, in which the word *probably* refers to a high probability, present the reasoning of NHST:

> If H_0 is true, then probably the test statistic will fall in the nonrejection region.
> The test statistic is in the rejection region.
> Therefore, H_0 is probably not true.

If the word *probably* is deleted from the first and third sentences, you get a valid modus tollens inference. However, when the word *probably* is retained, the inference is invalid, as demonstrated in the following example (Falk, 1986; Falk & Greenbaum, 1995).

For young women of age 30, the incidence of live-born infants with Down's syndrome is 1 in 885, whereas the majority of their pregnancies are normal. The amniocentesis test makes a positive diagnosis (+) in 99.5% of the Down's syndrome cases, and it has the same rate of correct results (–) in the case of normal pregnancies (denoted H_0). It can easily be verified, using Bayes's theorem, that if a young pregnant woman gets a positive test result, although this result is considered "significant" because $P(+|H_0) = .005$, the posterior probability of interest, namely that of normality, is $P(H_0|+) = .82$. Rejecting H_0 (diagnosing Down's syndrome) by NHST's prescription would evidently be a grave mistake (which may have undesirable consequences) in the face of a probability of over .80 that the fetus is normal. Cohen's (1994) example of a screening test for schizophrenia presents an isomorphic case (see also Pollard & Richardson, 1987). Thus, an improbable consequent does not necessarily render the antecedent improbable.

"The illusion of attaining improbability"(Falk & Greenbaum, 1995, p. 78), that is, the faulty belief that a statistically significant result makes H_0 improbable and deserving of rejection, is central to the NHST reasoning. If this belief is erroneous, the whole structure collapses: rejecting a hypothesis whose posterior probability is moderate or high is unacceptable. It is therefore odd that Hagen (1997) concluded that "the logic underlying statistical significance testing has not yet been successfully challenged" (p. 22). Moreover, he brought up puzzling examples to show that valid inferences are a luxury in scientific reasoning because they may result in unsound conclusions:

> If you contract AIDS, you will be healthy and happy.
> You did contract AIDS.
> You are healthy and happy. (p. 21)

This argument, structured as modus ponens, is valid in the inference drawn from an untrue first premise. You don't have to be an academic expert in "formal logic" to understand that if you start with an absurd, you'll end up with another absurd, despite applying a valid inference. In contrast, the first premise of NHST—if H_0 is true, then the probability of an outcome in the nonrejection region is high—is true. What is at fault here is the inference that if you get an outcome in the rejection region, then H_0 becomes improbable. Hagen may or may not be right in claiming that "science has done well using arguments that are not logically valid" (p. 22), but science has also done well without using NHST. The trouble with NHST is that it assumes the appearance of inferential validity and it may easily lead us astray.

There is indeed a lot of intellectual appeal to NHST (Falk, 1986; Falk & Greenbaum, 1995). Hagen's (1997) assessment that it will be hard to divorce ourselves from this practice is apparently right. If Fisher's (1960) tea-tasting-lady problem is posed to intelligent students, they immediately proceed to compute the probability of correctly identifying all eight cups, given that one is guessing (namely, H_0). They are also sure to interpret their computed low probability as that of guessing, given the perfect performance. These two conditional probabilities might differ considerably, however, depending on one's prior beliefs concerning the lady's ability (Lindley, 1993). The illusion of attaining improbability is strongly compelling, perhaps because it seems to satisfy our justified need to cope with chance.

REFERENCES

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

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98.

Fisher, R. A. (1960). *The design of experiments* (7th ed.). Edinburgh, Scotland: Oliver & Boyd.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15,* 22–25.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10,* 159–163.

Correspondence concerning this comment should be addressed to Ruma Falk, Department of Psychology, Hebrew University, Jerusalem 91905, Israel. Electronic mail may be sent to rfalk@cc.huji.ac.il.

In Praise of Brilliance: Where That Praise Really Belongs

Bruce Thompson
Texas A&M University and Baylor College of Medicine

Hagen (January 1997) offered "major points of disagreement" (p. 15) with Cohen's (1994) original work and urged celebration of the brilliance of the null hypothesis statistical significance test. Hagen simply ignored some of the persuasive warrants for the generally incisive conclusions presented by Cohen and, in other instances, seemingly misrepresented certain views offered by Cohen and other critics of conventional practice (cf. Kirk, 1996; Thompson, 1996, 1997), even though these other works were cited.

The present comment does not address the Bayesian and other mathematical arguments raised in both Cohen's (1994) original work and Hagen's (1997) criticism. In my view, these arguments are peripheral to Cohen's major points. Using mathematical arguments obfuscates understanding of the primary problems with conventional uses of statistical tests, and the argument can be fully joined and resolved without any resort to esoterica.

Regarding points ignored by Hagen (1997), he completely ignored one very major issue raised by Cohen (1994) involving what Cohen called "nil" (as against non-nil) null hypothesis testing. Most researchers mindlessly test only nulls of no difference or of no relationship because most statistical packages only test such hypotheses. This use of what Cohen called nil hypotheses does not require researchers to thoughtfully extrapolate expected results from the previous literature or from theory. Instead, science becomes an automated, blind search for mindless tabular asterisks using thoughtless hypotheses.

Cohen (1994) and others have taken the view that statistical tests would be more meaningful if more meaningful null hypotheses were used. Notwithstanding software impediments, for many hypotheses, researchers can evaluate meaningful parameters within statistical tests. This view merited consideration by Hagen (1997).

In addition, Hagen (1997) apparently misrepresented three critical points advanced by Cohen (1994). First, Hagen misrepresented Cohen's (and others') concerns about how the null hypothesis is used in statistical tests. Hagen argued at length that the null hypothesis is a statement about the population rather than the sample. But the issue is not so much about where the null hypothesis is assumed to be true; the concern involves how the null is used in statistical tests.

Statistical tests require that the null is assumed to be an exact, perfect description of truth in the population. This is required in order to have a single fully determined answer to the question "What is the probability of the sample statistics?" (Thompson, 1996).

Thus, statistical significance tests directly evaluate the probability of the sample statistics and do not directly evaluate the probability that the sample results also occur in the population. We as psychologists want to know about the population if we want to know whether our results will generalize and replicate; instead, we assume population parameters and test the sample. Cohen (1994) was absolutely correct in arguing that statistical testing really "does not tell us what we want to know" (p. 997).

Second, Hagen (1997) misrepresented Cohen's (1994) explanation as to why statistical significance tests are tautological. The null hypothesis is always false in the sample. Hagen struggled with why this is so; it's simply because the probability of any single point in a continuum of infinitely many sample statistics is itself infinitely small.

The consequence of the fact that the null is not exactly true in the sample (and I also don't believe populations exist where the "nil" of no difference is exactly true) means that the null will always be rejected at some sample size. Even the *Publication Manual of the American Psychological Association* (American Psychological Association, 1994, p. 18) recognizes that sample size largely drives rejection of the null hypothesis and therefore recommends reports of effect sizes; numerous empirical studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "recommending" has not appreciably affected reporting practices (e.g., Kirk, 1996; Thompson & Snyder, in press).

Thus, statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we've been too lazy to drag in enough participants.

As a discipline, we have been taken to the reductio ad absurdum of conducting power analyses to find the Goldilocks sample size that's "just right" enough to rescue statistical tests from being ridiculous (Levin, 1997). We want magical sample sizes somehow big enough to yield statistical significance but not so big as to stack the deck too much in favor of rejection when effects are ridiculously small.

Third, Hagen (1997) misrepresented the basis for recommending the use of confidence intervals. Hagen was conditionally correct in arguing that confidence intervals can invoke the same logic as statistical tests. The conditionality involves the question of how confidence intervals are interpreted.

If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing. But if the confidence intervals in a study are interpreted in the context of the intervals in all related previous studies, the true population parameters will eventually be estimated across studies, even if our prior expectations regarding the parameters are wildly wrong (Schmidt, 1996).

In summary, Hagen (1997) argued that "it is unlikely that we will ever be able to divorce ourselves from that [statistical test] logic even if someday we decide that we want to" (p. 22). However, notwithstanding this representation, we can alter our own behaviors if we deem such changes prudent and wise. If our minds decide that statistical tests do not evaluate either result importance (Kirk, 1996) or result replicability (Thompson, 1996), we really can tell our hands to type information that is relevant to these two important concerns. As we move toward more thoughtful inquiry, let's not forget to celebrate the brilliance of some of the perceptive criticisms of contemporary statistical practices and the brilliance of psychologists, such as the late Jack Cohen, who have pushed all of us to be more reflective.

REFERENCES

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

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

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12,* 84–106.

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

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29–32.

Thompson, B., & Snyder, P.A. (in press). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development.*

Bruce Thompson's World Wide Web address is http://acs.tamu.edu/~bbt6147/.

Correspondence concerning this comment should be addressed to Bruce Thompson, Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225.

Model Fitting: A Better Approach

Michael M. Granaas
University of South Dakota

Hagen's (January 1997) defense of the logic and practice of null hypothesis statistical testing (NHST) in response to Cohen's (1994) criticism is informative and troubling—troubling in that something so central to the practice of scientific psychology can be so difficult to understand and can require so much discussion. There simply has to be a better way.

Norman's (1993) book *Things That Make Us Smart* provides many examples of cognitive tasks that are difficult to perform given one approach to the task but that become much easier to perform given a formulation of the task more appropriate to the cognitive needs and structures of the person performing the task. NHST is an example of a cognitive task that is difficult to perform and error prone. Model fitting provides an approach to data analysis that is more appropriate to the cognitive needs of the researcher than is NHST.

Model fitting combines the NHST ability to falsify hypotheses with the parameter-estimation characteristic of confidence intervals in an approach that is simpler to learn, understand, and use. Effect size estimation is central to the approach, and power calculations are vastly simplified relative to NHST. Although model fitting is not perfect, it is better than the alternatives.

Using notation from Judd and McClelland (1989), model fitting involves comparing a reduced or compact model (Model C) with a full or augmented model (Model A) to determine which best represents, or fits, the data. In the simplest case, the compact model specifies a value for the parameter being estimated. The augmented model improves on the compact model by estimating the same parameter from the data. The proportional reduction in error (PRE) for Model A relative to Model C is evaluated to determine if the augmented model is better than the compact model. If the augmented model is better, Model A replaces Model C and becomes the new best estimate of the parameter. Rather than repeatedly testing null hypotheses of no difference and waiting for the meta-analyst to determine a parameter value, model fitting always has a current best estimate of the parameter value in place. This is exactly the strong form of NHST endorsed by Cohen (1994).

Consider two hypothetical independent researchers, Smith and Jones. Smith performs a one-sample *t* test by first declaring a null (nil) hypothesis value for some population mean (e.g., H_0: the population mean IQ for college graduates equals 100). The alternative hypothesis would be that the null hypothesis value is wrong. After collecting some data, Smith would presumably reject the value of 100 in favor of some unspecified other value. After many replications of this research, some significant, some not, a meta-analyst would compile the results of the replications to determine the correct value of IQ for college graduates.

Jones, using a model-fitting approach, would initially declare a Model C substantively identical to the H_0 used by Smith. After collecting some data, Jones would use the data to construct a competing model (e.g., the mean IQ for college graduates equals 108) and see which fits the data better. Like Smith, Jones would presumably reject Model C. Unlike Smith, Jones would declare 108 as the best estimate of the population mean. In replicating this research, 108 would become the new Model C (null) value to be either retained or replaced by a better estimate. After many replications of this research, the mean IQ for college graduates will have been estimated to some acceptable degree of precision.

With NHST, students are expected to learn an often confusing collection of techniques (regression, multiple regression, one-sample *t* test, independent groups *t* test, matched-pairs *t* test, and several variations of analysis of variance) with different hypotheses and applications. Model fitting uses a single computational technique, the general linear model, to replace the techniques listed. This makes model fitting easier to learn than NHST.

Model fitting, like confidence intervals, focuses on estimating parameters. The compact model provides the current best estimate of the parameter in question. If the augmented model fits the data better, it replaces the compact model, becoming the new compact model. Otherwise, the original compact model is retained. A best estimate of the parameter is always available.

Unlike confidence intervals, model fitting provides a means of falsifying parameter estimates. If two confidence-interval estimates of a parameter differ, there are no criteria for deciding which is best. With model fitting, there are criteria for choosing between the competing parameter estimates provided by Models A and C (i.e., PRE, which can be converted to an *F* value). Therefore, there is always a unique best estimate of a given parameter.

The estimation of power is simplified by consistent use of PRE and one-degree-of-freedom tests. One formula converts the sample value of PRE to an estimate of eta squared (the population effect size). A simple table lookup provides an estimate of power or the sample size needed to achieve a desired value of power.

The underlying mathematical basis for NHST, confidence intervals, and model fitting are identical and can yield the same conclusions. However, model fitting is superior in that its formulation makes it more appropriate to the cognitive structure and needs of the researcher. It is easier to learn and is less error prone than NHST. It is superior to evaluating confidence intervals in that it has the same goal—parameter estimation—embedded in a structure that supports falsification.

REFERENCES

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

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Norman, D. A. (1993). *Things that make us smart.* Reading, MA: Addison-Wesley.

Correspondence concerning this comment should be addressed to Michael M. Granaas, Heimstra Research Laboratories, Department of Psychology, University of South Dakota, 414 East Clark Street, Vermillion, SD 57069. Electronic mail may be sent to mgranaas@usd.edu.

A Further Look at Wrong Reasons to Abandon Statistical Testing

Richard L. Hagen
Florida State University

I am grateful to those who commented on my article (Hagen, January 1997). Their comments have helped me rethink and better understand some of the controversies surrounding the null hypothesis statistical test (NHST).

Critiques and criticisms of NHST over the past 40 years have tended to fall under three general topics: (a) the logical foundations of NHST, (b) the interpretation of NHST, and (c) alternative and supplementary methods of inference. The comments by Tryon (1998, this issue), McGrath (1998, this issue), Malgady (1998, this issue), Falk (1998, this issue), Thompson (1998, this issue), and Granaas (1998, this issue) also touch on all three of these topics, but the bulk of my response, as in the article (Hagen, 1997) to which the comments are directed, focuses on the first—namely, the logical foundations of NHST.

The logic of NHST has been challenged by three claims: (a) The null hypothesis is always false; therefore, a test of the null hypothesis is only a search for what is already known to be true; (b) the form of logic on which NHST rests is flawed; and (c) NHST does not tell us what we want to know. In attempting to rebut these claims, my position was, and still is, that "although there may be good reasons to give up the NHST, these particular points . . . are not among those reasons" (Hagen, 1997, p. 15). Several of the comments to which I am now responding again affirm (a) and (b), so that is where I begin.

Claim: The Null Hypothesis Is Always False

Thompson (1998) maintains that the null hypothesis is always false. If he is correct, then statistical testing would be, as he suggests, no more than a tautological search to find out what is already known. Let me respond to two points that he makes. First, Thompson states that the null hypothesis is always false in the sample. My response is that the null hypothesis is not a statement about the sample; therefore, the null cannot be false in the sample. The null hypothesis is a statement about the population from which the sample is drawn. Samples will always differ in an absolute sense if the measure is fine enough, but NHST anticipates such differences and accommodates them within the span of "$1 - \alpha$" in the sampling distribution of the statistic used to test the null.

Second, Thompson (1998) states that he does not believe populations exist where the "nil" of no difference is exactly true. I previously listed four arguments that have been used to support this belief that the null hypothesis is always false, and I attempted to refute, perhaps unsuccessfully—the reader has to decide—each of these arguments (Hagen, 1997, pp. 19–21). Space limitations do not permit a reiteration of these refutations. Nevertheless, the belief is apparently pervasive enough to merit an additional comment. I offer a rebuttal that I deleted from my 1997 article because reviewers felt it was overkill. The reviewers were correct. The thrust of that article would have been diminished if I had not removed that section. I offer it here as a last-ditch defense against the charge that the null hypothesis is always false.

If the null hypothesis is always false, then everything would have to be related to everything else. Just two variables in the universe that do not correlate would provide an example of the null hypothesis being true (H_0: the correlation between X and $Y = 0$). A correlation greater than zero requires some order; yet, it appears that disorder is more often expected than is order: "The second law of thermodynamics results from the fact that there are always many more disordered states than there are ordered ones" (Hawking, 1990, p. 145).

If the null hypothesis is always false, then all measurable human characteristics—indeed, temperament, intelligence, health, and even age at marriage and length of life—would have to be related, at least to some degree, to the position of the planets when one was born and the distribution of leaves in one's teacup. Voodoo rituals in Haiti would be related to rainfall in Montana, and social intelligence would be related to astrological signs. Scientists would even have to admit that the funding of their grant applications might be related to something other than the phases of the moon.

A reviewer of these comments wrote, "It is an unfair argument because none of the critics of the NHST have said that everything is related to everything else in a strong cosmic sense. Their contention is that δ and ρ are hardly ever zero." I suspect that this reviewer is correct—that critics have not meant that the null hypothesis is always false in a cosmic sense and that one should understand their message to be that the null hypothesis is almost always false, particularly in the lab. I must point out that if this is true, those same critics would have to modify or abandon the claim that NHST is no more than a search for what is already known to be true.

Claim: The Form of Logic on Which the Null Hypothesis Statistical Test Rests Is Flawed

Falk (1998) raises two criticisms of NHST, both of which challenge the logic of NHST. First, he restates the argument presented by Cohen (1994) that as a probabilistic imitation of modus tollens, NHST lacks formal logical validity. In my analysis (Hagen, 1997) of Cohen's article, I attempted to show in two ways that formal validity has little or nothing in common with reasonable scientific argument. First, I presented a clearly absurd inference that does have formal validity (Falk notes the absurdity of the inference and apparently agrees that it remains formally valid). Second, I attempted to demonstrate that arguments can be reasonable and defensible even

when they are not logically valid in a formal sense. My intent was to convince the reader that formal validity cannot represent a criterion against which any form of inference should be measured. I will not go over that ground again, but I attempt, in a somewhat different way, to defend the logic that underlies NHST. My effort is to show that this form of logic is accepted outside of the lab and that, therefore, scientists are on shaky ground if they deny its usefulness in the lab.

Suppose that an expert testifying in a murder case states that the blood of the defendant matches a sample found at the crime scene. Does a "match" mean anything? Well, it depends. A match on blood type means little or nothing, especially if the blood type is O. If the blood type is AB, a match may carry a little weight in the minds of some individuals but certainly not enough to push one to the very high standard of "beyond reasonable doubt." But a DNA match is a different matter. All are "matches." One is more weighty than another only because of the different probabilities of matches by chance alone if the defendant is, indeed, not guilty.

The apparent paradox of this form of reasoning is that the very information needed to provide evidence of guilt can have meaning only if the defendant is not guilty. As backward as this may seem at first glance, the argument is found to be compelling when it is applied to matches on DNA. Furthermore, according to Berger and Berry (1988), this form of logic is not unusual; rather, it is a familiar mathematical strategy known as "proof by contradiction."

The null hypothesis is as follows: The blood at the crime scene is not that of the defendant. If information is found that is very unlikely under this hypothesis, then it is rejected, and the only alternative is accepted: The blood at the crime scene is that of the defendant. A "match" on type O has a probability of about .46 of occurring by chance alone; that is not a small enough probability to reject the null. A "match" on type AB has a probability of occurring by chance alone of about .04, still not unusual enough to reject the null and convict someone of a crime. A match on DNA, however, as everyone has been told, has a probability of occurring by chance alone of less than one in many millions. And so the unlikely result of a match on DNA is considered strong evidence to knock down the null hypothesis.

As compelling as this example may appear to be, this form of reasoning is rarely invoked and therefore may not be easily grasped. Bruce Weir, a statistician–geneticist, testified at the O. J. Simpson trial that the probability of DNA matches by chance among the various stains tested was "one in 57 billion." Later, however, he questioned the wisdom of bringing the notion of chance into the courtroom:

> I was the expert witness called by the prosecution to tell the jury that these astronomical numbers are based on good scientific arguments. It would have been a lot easier if prosecutors could have simply said that forensics experts had found DNA matches. A radical concept? Not really. After all, fingerprint experts can testify that the defendant's prints were found at the crime scene, and no numbers are required to support their conclusion. (Weir, 1995, p. 11A)

Note that Weir (1995) recommended simply stating "there was a match," not because the argument by contradiction is flawed, but because of the difficulty most people have in tracking this kind of logic. Note also Weir's reference to testimony regarding fingerprints. The probability of a chance match of a fingerprint can vary depending on the matching number of characteristics (often based on the clarity of the print) and relative positions of those characteristics (Olsen, 1978). Yet, as Weir pointed out, in court, fingerprint experts can simply say they found a match. No statistics about the improbability of a match by chance are asked for or given. The same thing will probably happen in the future when experts testify about DNA matches.

I stand by my statement that the logic underlying NHST has not been successfully challenged. Examples from everyday life are infrequent, but they are there, and in terms of fingerprint and DNA matches, the logic is very compelling.

But before moving away from the logic of NHST, I want to consider one further criticism of this logic by Falk (1998), who mentions Cohen's (1994) "screening test for schizophrenia" example as having demonstrated that a significant result does not make H_0 improbable and deserving of rejection. Falk also states that I (Hagen, 1997) misrepresented the thrust of this example. He may be right in this regard.

I completely missed the possibility that in Cohen's (1994, pp. 998–999) example, a case (schizophrenic or normal) might be thought of as an entire experiment rather than a sample from a population about which an inference is to be made (this possibility was pointed out to me in an E-mail message from Nick Prins, a graduate student at the University of Kansas). If this was Cohen's intent, then I did misrepresent the thrust of his example.

With this frame of reference, Cohen's (1994) example can be applied to some population of experiments, 98% of which are "null hypothesis true" and 2% of which are "null hypothesis false." For all of the experiments, alpha is set at .03, and for the 2% "null false" experiments, beta equals .05. With this interpretation of Cohen's example, let us now revisit Falk's (1998) comment that the example demonstrates that a significant result does not make H_0 improbable and deserving of rejection.

Improbable? No. Not with the alpha Cohen (1994) used in this example. Cohen's figures show that when one obtains significant results, about 60% of those significant results would be from "null true" experiments. H_0 clearly is not improbable.

Less probable? Yes. As mentioned above, when significant results are obtained, the probability that we have in our hands a "null true" experiment is .60; however, when nonsignificant results are obtained, that probability is very close to 1.0. Cohen's (1994, p. 999) numbers show that out of 950 "negative test" cases (nonsignificant results), one could expect that 949 would be true negatives ("null true" experiments). Thus, a significant result signals a decrease of about 40% in the probability of H_0.

But there is yet another important lesson about NHST that can be gained from Cohen's (1994) example, namely, how an adjustment in alpha can lead to a decrease in the probability that H_0 is true given significant data, perhaps to the point that H_0 is worthy of rejection.

Let me try to illustrate by referring again to Cohen's (1994, p. 999) example. Assume a population of 1,000 experiments on the beneficial effects of various foods, herbs, or tonics on specific health conditions (as compared with placebo controls). Ninety-eight percent of these "treatments" are ineffective; 2% are effective. By using the alpha and beta in Cohen's example, we can expect to correctly identify 19 of the 20 treatments that really do work. Some might say that's not bad sleuthing.

But we also would expect to make mistakes on about 30 of the ineffective treatments by calling them effective. Is it reasonable to act as if these treatments are effective when they are not? That question can be answered only within the framework of the costs and benefits associated with Type I and Type II errors (see Malgady, 1998). If a treatment is cheap, is easy to obtain, and has minimal side effects—for example, in the case of garlic or broccoli—Type I errors are not of great concern. But if the "treatment" is more expensive and has a few more side effects—like red wine—we may want to ratchet down alpha a bit to reduce Type I errors. If the "treatment" tastes bad or has serious side effects, we would want to tighten alpha even more.

In the example being considered, just by reducing alpha from .03 to .01, the expected

misidentification of ineffective treatments (false positives) is reduced by 67% (from about 30 to about 10), whereas the expected correct identification of effective treatments (true positives) is reduced by only 10% (from about 19 to about 17). If the judgment is that false positives are considerably more costly than are false negatives (Type I errors are considerably more costly than are Type II errors), and alpha is accordingly reduced to .001, it would be expected that only 1 out of 980 ineffective treatments would be misidentified while 13 out of the 20 effective treatments would still be correctly identified. Not only is this overall expected hit rate quite impressive, but in addition with this stringent alpha, the probability that H_0 is true given significant data (the Bayesian inverse probability) is only .07. Further decreases in alpha, with the same effect size and a concurrent adjustment in beta, would render H_0 even more "improbable and [perhaps] deserving of rejection" (Falk, 1998, p. 798).

When Cohen's (1994) example is cast in the framework of experiments, rather than in cases of schizophrenic or normal individuals, it provides a good illustration of how NHST can assist researchers in appraising propositions. In spite of the great imbalance in base rates in this example and regardless of how stringently alpha is set, the probability that H_0 is true is always lower when significant results are obtained than it is when significant results are not obtained. Thus, NHST does provide information about the probability that H_0 is true, which, after all, is what "we so much want to know" (Cohen, 1994, p. 997).

Closing Remarks

I have only a few more remarks about the preceding comments, and these remarks relate only peripherally to my analysis (Hagen, 1997) of Cohen's (1994) article. First, Malgady (1998) draws our attention to the problems associated with "accepting the null hypothesis." My only caution—a minor caution—about Malgady's comment is that we not apply too broadly his statement that when we behave as if the null hypothesis is true, we validate it or accept it. Sometimes this may be true; sometimes it may not be.

I recently encountered a smoker, a researcher, and asked him if he believes smoking is harmful to one's health. He answered, "Absolutely. And I will probably die of some condition that will result from what I am doing right now." I then asked, "As one acquainted with research methods, do you think you are validating or affirming a null hypothesis about smoking and health?" He looked at me with a wry smile and replied, "Certainly not. My choice to smoke has nothing to do with research."

With this caveat, I applaud Malgady's (1998) reminder that the classical Fisherian procedure does not allow one to accept or validate the null hypothesis. To do so can lead not only to misrepresentations of what the procedure can and cannot do but also to errors in interpreting results. For example, accepting the null hypothesis may seduce us into a "meta-analysis by tally," and we may later be embarrassed when a modern meta-analysis reminds us that *absence of evidence* does not equal *evidence of absence*.

Second, Thompson (1998) mentions that I (Hagen, 1997) ignored the issue of "nil" hypothesis testing. That is an issue related to use, or misuse, of NHST, not one related to the integrity of the method itself. Even Cohen (1994) said that for some questions, "any departure from pure chance is meaningful" (p. 1000). It is up to the researcher to judge when that might be true.

Third, would statistical tests be more meaningful if more researchers were more thoughtful in the way they use statistical inference (McGrath, 1998; Thompson, 1998)? Without a doubt.

Fourth, should effect sizes and confidence intervals always be reported (Thompson, 1998)? By all means.

Fifth, McGrath (1998) notes that I (Hagen, 1997) did not offer a strong case for the continued use of NHST as a primary inferential strategy in psychology. I am glad he pointed this out because some readers may have thought that I was suggesting that NHST should be the primary inferential strategy in psychology. I am struck by the beauty, elegance, and usefulness of NHST, but other methods of inference may be equally elegant and even more useful depending on the question being asked. Hopefully, better and better inferential strategies will continue to be developed. Granaas (1998) suggests that model fitting is superior to both confidence intervals and NHST. He may be right. What little I know about model fitting tells me that it is a very useful form of inference, but at this point in my development, I am not competent to judge model fitting relative to other forms of inference.

Sixth, Tryon (1998) points out that there are substantial reasons to seriously question whether NHST results have been, are, or can be correctly interpreted consistently by most investigators. And McGrath (1998) suggests that maybe NHST should be given up simply because it is so frequently misunderstood and misinterpreted. Tryon is certainly right, and alas, McGrath may be right also. But I remind the reader that the misunderstandings and misinterpretations are our problem, not the problem of NHST.

This interchange has been rewarding to me. I hope it has been for the reader. A continuing dialogue may, indeed, lead our field to give up NHST in favor of other methods of inference. If that happens, I hope this is done through understanding, rather than misunderstanding, the strengths and limitations of NHST.

REFERENCES

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

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

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53,* 798–799.

Granaas, M. M. (1998). Model fitting: A better approach. *American Psychologist, 53,* 800–801.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Hawking, S. (1990). *A brief history of time: From the big bang to black holes.* New York: Bantam Books.

Malgady R. G. (1998). In praise of value judgments in null hypothesis testing . . . and of "accepting" the null hypothesis. *American Psychologist, 53,* 797–798.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53,* 796–797.

Olsen, R. D. (1978). *Scott's fingerprint mechanics.* Springfield, IL: Charles C Thomas.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799–800.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53,* 796.

Weir, B. S. (1995, October 3). One in 57 billion: The use of DNA forensics is on trial as well as O. J. *Charlotte Observer* [Special to the Observer Viewpoint Section], p. 11A.

Correspondence concerning this comment should be addressed to Richard L. Hagen, Department of Psychology, Florida State University, Tallahassee, FL 32306-1051. Electronic mail may be sent to hagen@psy.fsu.edu.

In Defense of Deception

Allan J. Kimmel
Ecole Supérieure de Commerce de Paris

Ortmann and Hertwig's (July 1997) recent call to outlaw the use of all forms of deception in psychological research is, in my view, both methodologically unsound and ethically misguided. This, of course, is not the first time that researchers have voiced their concerns about the potentially ill effects of

deceptive research tactics in human participant research. Critics of deception have long decried its use in psychology on the basis of strictly moralistic grounds (e.g., regardless of the anticipated research ends, it is always wrong to mislead research participants), methodological grounds (e.g., deception increases future research participants' suspiciousness), or more general disciplinary considerations (e.g., deception reduces trust in psychologists and gives the profession a poor reputation).

Whatever the specific complaint, the criticisms leveled against deception, in one form or another, suggest that because it involves lying and deceit, its use in psychological research is morally reprehensible and may have potentially negative effects on participants, the profession, and society. Basing their arguments on the logic of game theory, Ortmann and Hertwig (1997) essentially revisited familiar territory by focusing on two potential drawbacks to an unbridled use of deception in research: (a) that research participants' behavior will be affected in unintended ways by the expectation that they will be misled in the research context and (b) that the profession of psychology will experience "reputational spillover effects" (p. 747) as mass-mediated accounts of research increasingly reveal the deceptive tactics used by a growing number of psychologists.

Generally speaking, one cannot take exception to Ortmann and Hertwig's (1997) central claim that the prevalence of deception has risen over the years, given that this has been well-documented in several surveys of methodological and ethical practices in psychology (e.g., Gross & Fleming, 1982). It is important to note, however, that these surveys are now dated and they shed little light on what has been going on in psychology during the past 15 to 20 years. Ortmann and Hertwig implied that the frequency of deception continues to rise, citing a secondary source account of an analysis conducted by Adair, Dushenko, and Lindsay (1985) to support their claim. However, Adair et al.'s analysis focused on empirical studies appearing in psychology journals during 1979 and 1983. Since that time, a formidable array of ethical guidelines and review mechanisms have evolved. It is likely that in recent years the frequency of deception has at least leveled off and perhaps even declined. The kind and degree of deception also may have changed in recent decades, with researchers relying more on deceptions of the passive sort (e.g., withholding relevant information from participants) than the active variety (e.g., blatantly misleading participants). No doubt the era of the "fun and games approach" taken by many psychologists in their attempts to create increasingly elaborate deceptions has ended. What is clearly needed is an up-to-date assessment of the nature and frequency of deceptive techniques in psychological research.

Some critics of deception claim that any amount or kind of deception in research could ultimately have adverse effects on the behavior of research participants and the profession's reputation once its use becomes common knowledge. Contrary to the fears expressed by Ortmann and Hertwig (1997), although prospective research participants have long been aware of the possibility of being deceived in psychology experiments, they generally have remained cooperative. Since the 1960s, researchers have warned that participants in psychology experiments have considerable awareness of the implicit rules that govern the situation and have grown to distrust experimenters because they know that the true purpose of the experiments may be disguised. By the early 1970s, investigators reported high rates of suspiciousness among participants in conformity studies, ranging from 50% to nearly 90% (e.g., Glinski, Glinski, & Slatin, 1970). Nonetheless, the effects of suspiciousness on research performance, though somewhat inconsistent, appear to be negligible, leading some to conclude that, in general, there are not major differences between the data of suspicious and reportedly naive participants (Kimmel, 1996). When effects have been found, they have resulted in participants' tendency for favorable self-presentation (such as improving their performance on problem-solving tasks) rather than influencing their motivation to cooperate.

Aside from the minimal effects of prior suspiciousness on research performance, the preponderance of evidence suggests that deceived participants do not become resentful about having been fooled by researchers and that deception does not negatively influence their perceptions about psychology or their attitudes about science in general (Kimmel, 1996). For example, in a review of studies that assessed research participants' reactions to deception experiments, Christensen (1988) concluded that persons who have participated in deception (vs. nondeception) experiments reported that they did not mind being deceived, enjoyed the experience more, received more educational benefit from it, and did not perceive that their privacy had been invaded. Furthermore, the results of surveys intended to gauge reactions to deception have consistently shown that most individuals in the general population apparently do not have serious objections to its use for research purposes. There also is evidence that psychologists have more serious reservations about the use of deception than do university students—the very persons who comprise the typical research population and who are likely to experience harm from its use. More recently, Sharpe, Adair, and Roese (1992) revealed that there has not been a predicted increase in negative attitudes toward psychological research in the student population as a result of the continued use of deception during the past 20 years.

To support their contention that deception eventually will result in reputational spillover effects, Ortmann and Hertwig (1997) reproduced a brief item from *The International Herald Tribune* describing the multiple active deceptions utilized in a cross-cultural comparison of stress and aggression. Although the example can hardly be said to present psychologists in a favorable light, it is clearly the exception rather than the rule. More typically, media accounts of psychological research tend to focus on findings of particular interest to the public (on such inherently interesting topics as shyness, helping behavior, rumor and gossip, and gender differences); rarely is much ever said about the methodology used. Here is an example, in its entirety: "A psychology researcher in Chicago cured stuttering in an 8-year-old girl and helped three other children by catching their fluent speech on videotape and using the scenes as a model" ("A Psychology Researcher," 1997, p. 3). In my experience, it is just these sorts of reports that often attract new students to the discipline of psychology. Moreover, as Rosnow (1997) pointed out, the proliferation and increased role of ethics committees, legalities, and other external restrictions have already subjected psychologists to a higher level of professional ethical accountability than is found in many other professions (including law, politics, and marketing), where both passive and active forms of deception are commonplace.

To be sure, I am not recommending that deception be used as a matter of course by psychologists. However, deception procedures differ so much in the nature and degree of deception used that even the harshest critic would be hard-pressed to state unequivocally that all deception is unacceptable. When the methodological requirements of an investigation lead the researcher to conclude that the only way a study can be carried out is by using deceptive research tactics, the decision to deceive necessarily results in additional ethical responsibilities for the researcher, and the degree of deception should be held at a level that does not exceed what is required by the research (American Psychological Association, 1992).

Realistically, ethical research procedures such as informed consent are not always the most methodologically sound procedures. In other words, what is the most ethical is not necessarily the most effective, and the potential loss of important research benefits from the decision not to do a study needs to be weighed as seriously as the risks involved in

doing the study. An absolute rule prohibiting the use of deception in all psychological research would have the egregious consequence of preventing researchers from carrying out a wide range of important studies.

REFERENCES

Adair, J. G., Dushenko, T. W., & Lindsay, R. C. L. (1985). Ethical regulations and their impact on research practice. *American Psychologist, 40,* 59–72.

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

Christensen, L. (1988). Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin, 14,* 664–675.

Glinski, R. J., Glinski, B. C., & Slatin, P. T. (1970). Nonnaivety contamination in conformity experiments: Sources, effects, and implications for control. *Journal of Personality and Social Psychology, 16,* 478–485.

Gross, A. E., & Fleming, I. (1982). Twenty years of deception in social psychology. *Personality and Social Psychology Bulletin, 8,* 402–408.

Kimmel, A. J. (1996). *Ethical issues in behavioral research: A survey.* Cambridge, MA: Blackwell.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist, 52,* 746–747.

A psychology researcher in Chicago. (1997, August 19). *The International Herald Tribune,* p. 3.

Rosnow, R. L. (1997). Hedgehogs, foxes, and the evolving social contract in psychological science: Ethical challenges and methodological opportunities. *Psychological Methods, 2,* 345–356.

Sharpe, D., Adair, J. G., & Roese, N. J. (1992). Twenty years of deception research: A decline in subjects' trust? *Personality and Social Psychology Bulletin, 18,* 585–590.

Correspondence concerning this comment should be addressed to Allan J. Kimmel, Département Marketing, Ecole Supérieure de Commerce de Paris, 79, avenue de la République, 75543 Paris Cedex 11, France. Electronic mail may be sent to kimmel@escp.fr.

The Reality of Deception

James H. Korn
Saint Louis University

Ortmann and Hertwig (July 1997) are concerned about a "dramatic increase since the early 1960s" (p. 747) in the use of deception in psychological research. Deception as a research technique has been a common practice in social psychology since the 1960s; and during the 1970s, there was an increase in deceptive research; but from then through 1994, there appears to have been a decrease (Korn, 1997; Nicks, Korn, & Mainieri, 1997). The practice still is relatively frequent, but instances of dramatic impact experiments that were common in the 1970s are exceptional today because of changes in theory, method, and ethical standards.

One ethical standard cited by Ortmann and Hertwig (1997) is the previous version of the ethical principles of the American Psychological Association (APA). The current version (APA, 1992) is more clear on this issue: "Psychologists *never* [italics added] deceive research participants about significant aspects that would affect their willingness to participate, such as physical risks, discomfort, or unpleasant emotional experiences" (p. 1609). All U.S. universities have Institutional Review Boards composed of nonpsychologists who oversee risk–benefit decisions based on ethical standards, including those of APA. The extent to which this oversight is carefully implemented, however, is cause for concern.

Social psychologists in particular are interested in research questions that often can be studied only if deception is used in realistic situations. The history of the use of deception in social psychology is linked not only to changes in psychological theory and methods but also to characteristics of American culture such as individualism and pragmatism (Korn, 1997). Thus, individual experimenters decide if the results of their research justify the use of deception. Moral philosophers do not agree that deception always is wrong, and in our cultural context, the suggestion by Ortmann and Hertwig (1997) that all forms of deception be outlawed is unrealistic.

REFERENCES

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

Korn, J. H. (1997). *Illusions of reality: A history of deception in social psychology.* Albany: State University of New York Press.

Nicks, S. D., Korn, J. H., & Mainieri, T. (1997). The rise and fall of deception in social psychology and personality research, 1921 to 1994. *Ethics and Behavior, 7,* 69–77.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist, 52,* 746–747.

Correspondence concerning this comment should be addressed to James H. Korn, Department of Psychology, Saint Louis University, 221 North Grand Boulevard, St. Louis, MO 63103. Electronic mail may be sent to kornjh@slu.edu.

Deception Can Be Acceptable

Arndt Bröder
University of Bonn

Ortmann and Hertwig (July 1997) vehemently argued against any deception of participants in psychological experiments. Like the authors, I think that ethical principles in psychological research are in fact not debated with the priority they deserve. However, some clarifying comments concerning Ortmann and Hertwig's claims are certainly necessary. Their sophisticated arguments against deception in research are neither precise nor imperative, addressing only questionable negative long-term effects of deception in research. In this comment, I argue that (a) acceptability of an experimental treatment and acceptability of deception must be kept separate, (b) deception is necessary in research on certain topics, and (c) participants understand and even accept deception when they are carefully debriefed. I begin with a short summary of Ortmann and Hertwig's arguments.

Ortmann and Hertwig's (1997) main point is that cooperative participants will become uncooperative if they are repeatedly deceived in psychological experiments. They drew on well-known results from the repeated prisoners' dilemma game, in which repeated noncooperation of one partner elicits noncooperation in the decisions of the other partner. As a consequence, if participants are repeatedly deceived, in the long run, psychology as a profession in general will get an increasingly bad image, causing participants to be suspicious and uncooperative in future research. For this reason, the American Psychological Association's ethical principles require a deliberate cost–benefit evaluation before participants are deceived. Unfortunately, according to Ortmann and Hertwig, this mechanism will not work because it is a trade-off between individual benefits and public costs having the same structure as a so-called social trap (e.g., Platt, 1973). If this self-monitoring mechanism does not work, any deception should be abandoned from research as it is done in experimental economics.

What Do Ortmann and Hertwig (1997) Mean by "Deception?"

Ortmann and Hertwig (1997) did not specify their use of the term *deception*; they merely illustrated it by citing a study by Cohen, Nisbett, Bowdle, and Schwartz (1996) in which participants indeed were bumped into and called "assholes" by a confederate of the experimenter in order to test their emotional

reactions. Of course, participants did not know this was part of the experiment and were deceived about the real experimental purpose in this respect. At first glance, this study seems to be an intriguing example of what could be called an unacceptable deception of participants. But further examination shows that two aspects are intertwined here that should be kept separate: What might be considered obnoxious in this way of treating volunteer participants is the treatment (calling them "assholes") rather than the act of deceiving. One can hardly imagine somebody expressing scrupulosity if the treatment had consisted of a friend passing by saying "hello" even if participants were deceived about the real purpose of the experiment in the same way (i.e., measuring emotional reactions). The acceptability of treatment and of deception about the purpose of an experiment are different things and must be evaluated separately with respect to ethical appropriateness. So Ortmann and Hertwig's claim for abandoning deception completely means "throwing the baby out with the bath water."

Is Deception Needed in Psychological Research?

Deception may be defined as concealing or camouflaging the real purpose of an experiment (i.e., the data in which the scientist is interested) to avoid conscious reactivity of participants that would make these data worthless. In fact, memory research in large areas would be impossible if this kind of deception was not allowed. Consider the research on incidental learning. Participants are told to rate stimuli on some emotional dimensions or to do some other (often irrelevant) task on them, certainly not knowing that a memory test will follow. If they were told about this fact in advance, it would not be a study on incidental learning by definition. Plausible, but necessarily deceptive, cover stories have to be used in these cases. In studies of cognitive illusions (e.g., hindsight bias or misleading postevent information effect), it is a necessity to conceal the true nature of the experiment. These are only two of numerous examples. The ethical question concerning deception in this research therefore cannot be whether deception is necessary within this research (because it is) but rather whether this research is necessary. This must of course be the topic of public discussion in which psychologists will have to defend their claims about the relevance of their research. But this is the case for every empirical science. It is of no help for cognitive psychologists when Ortmann and Hertwig (1997) noted that "in experimental economics, for example, professional conventions categorically prohibit deception" (p. 747) because deception may not be necessary in most studies on economic decisions. So economists easily can do without this tool, whereas psychologists often cannot.

Do Participants Become Uncooperative?

This last section is based on my experience as a participant as well as an experimenter. My impression is that most people participating in psychological experiments are very interested in the purpose of these studies. Psychological research results are relevant for almost everyone. If participants are carefully informed about the purpose of the experiment and the necessity of deception (e.g., in cognitive illusion research), most of them will accept this deception as an indispensable tool. This is reflected in the fact that most of the participants in studies conducted at our department agreed to participate again in other experiments even after having been debriefed and informed about the real purpose of the studies. Because they are volunteers, they easily could terminate the sequence that Ortmann and Hertwig (1997) called a "repeated prisoners' dilemma game." In fact, most of them do not withdraw. It should go without saying that participants must be debriefed about every experimental manipulation, including deception. If this is carefully done, I do not expect the dramatic image loss of psychology as a profession in general, which Ortmann and Hertwig expect.

Interestingly, Ortmann and Hertwig's (1997) line of argument is in no way ethical but purely pragmatic. Despite this fact, I agree with the authors about the importance of careful ethical considerations of any treatment in psychological research. As in any empirical science, the trade-off between possible harms of interventions (costs) and scientific relevance (benefit) should be a matter of public discussion. I would like to endorse that deception should be avoided whenever possible, but in some cases (e.g., incidental learning), this cannot be done without sacrificing the purpose of research. The most problematic point in Ortmann and Hertwig's arguments is their confounding of treatment and deception by simply citing one example that is not very typical for experimental psychology in general. By doing this, they evoke the image of psychological laboratories being places where Milgram studies are commonplace. Most psychologists would agree that this is far from the truth. Not clarifying the distinction between the acceptability of a treatment and the acceptability of deception, Ortmann and Hertwig might cause a greater (and undeserved) image loss of psychology than deception itself.

REFERENCES

Cohen, D., Nisbett, R., Bowdle, B. F., & Schwartz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology, 70,* 945–960.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist, 52,* 746–747.

Platt, J. (1973). Social traps. *American Psychologist, 28,* 641–651.

Correspondence concerning this comment should be addressed to Arndt Bröder, Department of Psychology, University of Bonn, Römerstrasse 164, D-53117 Bonn, Germany. Electronic mail may be sent to arndt.broeder@uni-bonn.de.

The Question Remains: Is Deception Acceptable?

Andreas Ortmann
Bowdoin College

Ralph Hertwig
Max Planck Institute for Human Development, Berlin

In response to our comment titled "Is Deception Acceptable?" (Ortmann & Hertwig, July 1997), Kimmel (1998, this issue) and Korn (1998, this issue) question our assertion that the use of deception in psychological experiments has increased since the early 1960s. Korn cites two of his own studies as showing that "during the 1970s, there was an increase in deceptive research; but from then through 1994, there appears to have been a decrease" (p. 805). His results conflict with those of Sieber, Iannuzzo, and Rodriguez (1995), who reported that in the top-ranking social psychology journal, *Journal of Personality and Social Psychology* (which Nicks, Korn, & Mainieri, 1997, also analyzed), the percentage of studies using deception has remained essentially the same since the 1970s, despite a dip in the mid-1980s (47% in 1978, 32% in 1986, and 47% in 1992). The discrepany between their results could stem from definitions of deception that differ in inclusiveness.

Whether there has been a decline in the number of studies using deception (by any definition) in recent decades, however, is irrelevant to our argument. Even if its use is less frequent and less dramatic than in the past, deception can strongly affect the reputation of individual labs and the profession,

thus contaminating the participant pool. If participants arrive at an experiment knowing that they may be deceived, distrusting the experimenter as a result, then control over the experimental conditions is compromised. The question is not whether one has less of a bad thing but whether one has a bad thing at all.

Of course, whether deception is a bad thing methodologically (never mind ethically) is a question open to dispute. We believe that deception significantly influences the behavior of participants, whereas Kimmel (1998), Bröder (1998, this issue), and others do not. Kimmel cites several studies that seem to suggest that participants have a positive attitude toward the use of deception in psychological experiments. Unfortunately, all of them measured participants' attitudes rather than their actual behavior. Even if one believes the finding in these studies that participants do not mind deception, one cannot therefore assume that they behave cooperatively in experiments in which they expect to be deceived. In fact, there is evidence that they do not (e.g., MacCoun & Kerr, 1987; Newberry, 1973; Taylor & Shepperd, 1996). Still, the question of whether deception matters deserves further inquiry.

In closing, we would like to note that our definition of deception does not coincide with that intimated by Bröder (1998). To us, not telling participants the purpose of an experiment is not necessarily deception; telling participants things that are not true necessarily is.

REFERENCES

Bröder, A. (1998). Deception can be acceptable. *American Psychologist, 53,* 805–806.

Kimmel, A. J. (1998). In defense of deception. *American Psychologist, 53,* 803–805.

Korn, J. H. (1998). The reality of deception. *American Psychologist, 53,* 805.

MacCoun, R. J., & Kerr, N. L. (1987). Suspicion in the psychological laboratory: Kelman's prophecy revisited. *American Psychologist, 42,* 199.

Newberry, B. H. (1973). Truth telling in subjects with information about experiments: Who is being deceived? *Journal of Personality and Social Psychology, 25,* 369–374.

Nicks, S. D., Korn, J. H., & Mainieri, T. (1997). The rise and fall of deception in social psychology and personality research, 1921 to 1994. *Ethics & Behavior, 7,* 69–77.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist, 52,* 746–747.

Sieber, J. E., Iannuzzo, R., & Rodriguez, B. (1995). Deception methods in psychology: Have they changed in 23 years? *Ethics & Behavior, 5,* 67–85.

Taylor, K. M., & Shepperd, J. A. (1996). Probing suspicion among participants in deception research. *American Psychologist, 51,* 886–887.

Correspondence concerning this comment should be addressed to Andreas Ortmann, Bowdoin College, Brunswick, ME 04011. Electronic mail may be sent to aortmann@bowdoin.edu.