Significance Tests are Not Enough: The Role of Effect-Size Estimation in Theory Corroboration

Monica J. Harris

Theory Psychology 1991; 1; 375
DOI: 10.1177/0959354391013007

COMMENT 3

Significance Tests Are Not Enough

The Role of Effect-size Estimation in Theory Corroboration

Monica J. Harris
UNIVERSITY OF KENTUCKY

ABSTRACT. Chow (1991) distinguishes between 'practical impact' and 'conceptual rigor' research, and he concludes that effect-size estimation is useful only in practical impact research. I argue that significance tests do not answer substantive questions about the data and are useful only as a check that the results are unlikely to have occurred by chance. Chow's decision to regard the similarity between data and prediction as being a dichotomous judgment made on the basis of significance testing is therefore unwise. I conclude that effect sizes are the single best index of the relationship between theoretical predictions and the obtained data. The role of replications and meta-analysis in advancing theory is also discussed.

Questioning the utility of significance tests is not new. There is a long and honorable tradition of blistering attacks on the role of significance testing in the behavioral sciences (e.g. Bakan, 1966; Grant, 1962; Hedges & Olkin, 1985; Lykken, 1968; Meehl, 1978; Rosenthal, 1984), a tradition reminiscent of knights in shining armor bravely marching off, one by one, to slay a rather large and stubborn dragon. Meehl's (1978) indictment of 'tabular asterisks' is a prime example of this tradition; he pulls no punches when he says:

> . . . the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology (p. 817).

Given the cogency, vehemence and repetition of such attacks, it is surprising to see that the dragon will not stay dead. Chow's article in this issue is the latest in a series (Chow, 1988, 1989, 1991) wherein he proposes that significance tests are an 'indispensable step' in the theory-corroboration

THEORY & PSYCHOLOGY © 1991 SAGE. VOL. 1(3): 375–382

process and that effect-size estimation is 'antithetical to the requirements of internal validity and of external validity' (p. 337). Chow's basic thesis is to divide psychological research into two domains: (a) research investigating the 'practical impact' of some experimental manipulation; and (b) research aimed at evaluating and advancing theory, called the 'conceptual rigor' perspective. He claims that effect-size estimation is necessary only for assessing the practical impact question, and that it is not a necessary or useful part of theory corroboration. I believe that Chow's conclusions are incorrect and that a renewed emphasis on significance testing would be damaging to the scientific enterprise of testing and refining theories. The goal of this commentary, then, is to provide a challenge to Chow's arguments and to discuss the proper roles of significance testing and effect-size estimation in theory corroboration in the behavioral sciences.

On the Limitations of Significance Tests in Theory Evaluation

The 'conceptual rigor' perspective in psychological research is described in Chow's Table 3, which outlines the steps involved in testing a conceptual hypothesis. To summarize, a researcher derives an implication of a theory, identifies an empirical prediction to test that implication, collects data and compares the data to the prediction. If the data are similar to the prediction, then the conclusion that the theoretical implication is probably correct is drawn. If the data are dissimilar to the prediction, the researcher concludes that the implication and theory are false. Significance tests are used to make a binary decision about the similarity of the data to predictions.

My major objections concern the following conclusions: (a) that the judgment of similarity between data and predictions is best considered a binary one; and (b) that estimates of effect size are not informative in making such judgments.

Judgments of Similarity Are Continuous, Not Qualitative

Chow's conclusions about the utility of significance testing rest on the assumption that similarity is best considered a dichotomous judgment. However, such an assumption is at best imprudent. Similarity is a continuous dimension. Of course, it is possible to force a dichotomous judgment of similarity by drawing an arbitrary line on the continuum between 'similar' and 'dissimilar'; and if one were to do so, the rest of Chow's argument follows logically. But merely because a practice is logically *possible* does not make it *desirable*. In the case of similarity, it is neither necessary nor desirable to draw the line at all.

A good analogy is seen in a frequent data-analytic practice: a researcher administers a continuous individual difference measure, say the WAIS,

and then promptly proceeds to perform a median split on the measure, divides the sample into two groups and then tests the difference between the groups on some dependent measure through a *t*-test or ANOVA. This procedure is certainly *logically permissible*, but it is just as certainly stupid. Not only does it result in an avoidable loss of power, but it also regards a subject with an IQ of 99, for example, as being qualitatively different from a subject with an IQ of 101. Analogously, using significance testing to decide whether data are similar to predictions could lead one to conclude—incorrectly—that a result that was 'nonsignificant' at $p = .053$ was qualitatively different from a result yielding $p = .049$.

I will now apply the same reasoning to an issue more relevant to Chow's arguments: choosing between rival explanatory hypotheses. Suppose we are investigating our favorite theory, and we postulate our favorite explanatory mechanism. We realize that there are potential rival explanatory mechanisms that we would like to be able to rule out. On the basis of these mechanisms we can form predictions about the results. If our favorite explanatory mechanism is right, the data should look like *X*. If the rival explanatory mechanism is right, the data should look like *Y*. We then conduct our study, analyze the data and ask ourselves 'which mechanism is right?' How can such a decision be made?

Chow would have us make that decision on the basis of significance tests. If the means are significantly different in the direction predicted by our theory, then we have support for our explanatory mechanism. However, such a tactic can lead us into serious difficulty. Suppose that the obtained data support *both* explanatory mechanisms significantly (as is the case, for example, when an S-shaped curve exhibits linear and quadratic trends that are both statistically significant). How then to evaluate the two competing theories? Chow's approach leaves us with no answer to that question. An effect-size approach would have us compute effect sizes for the contrasts predicted by each theory. We can then compare the two effect sizes to determine which theory does the better job of explaining the data. If we should desire a formal statistical comparison of the effect sizes, we can use Hotelling's *t*-test for the significance of the difference between two dependent correlations.

We would fail miserably in our goal of explaining behavior if we treated all statistically significant results as equal. One may try to argue, as Chow does, that all significant results imply similarity between theory and data, but a more reasonable statement is that, *à la* Orwell, some results are more similar than others. We would also fail miserably in our goal of explaining behavior if we regarded a p of .051 as being qualitatively different and *substantially worse than* a p of .049. The binary judgment afforded by significance tests does not yield an optimal solution to the question of similarity because similarity is a continuous variable, not a dichotomous variable.

Effect Sizes Are More Informative than Significance Tests

Chow and I are in agreement that the heart of theory corroboration is determining the degree of fit between one's theoretical implications and the obtained data. However, we differ in that I believe that effect sizes are more useful in answering this question than are significance tests. When we compute a contrast, we are essentially computing the fit between the obtained data and a focused hypothesis expressed by the chosen contrast weights (Rosenthal & Rosnow, 1985). The effect size associated with the contrast represents the *single most informative index* of the relationship between our theory and the real world.

I do not mean to imply that significance tests are unimportant; they provide a useful check that the results of our contrasts are not due merely to random sampling fluctuation. However, that is all they do. They do not tell us *how well* our theory captures the true state of affairs, a question that is addressed by effect-size estimation. For example, it is informative to say that 'although the predicted difference is statistically significant, the effect is very small'. Such a statement tells us that we have a long way to go in our quest for explaining the behavior in question, a fact of considerable importance to researchers in refining their theories and planning future research. Much is also gained by a statement such as 'although both contrasts are statistically significant, the contrast predicted by Theory A has a substantially larger effect size than that of the contrast predicted by Theory B'. A careful comparison of the magnitudes of effect sizes of various contrasts allows us to make a more precise evaluation of competing theories.

Power and Type II Errors: Limitations of Significance Tests

Chow acknowledges, and then dismisses, the argument that an emphasis on significance testing neglects the likelihood and seriousness of committing Type II errors. He contends that researchers weigh the relative importance of Type I and Type II errors through the 'judicious choice of the alpha value' (p. 346). However, we do not currently possess the flexibility to choose judiciously the alpha level we would like to employ; levels more liberal than .05 or, less often, .10 are not accepted by the scientific community.

Chow is correct when he says that significance tests should not be blamed when researchers misuse them, and as responsible scientists we should strive to obtain as many subjects as needed to yield a fair test of the hypothesis. However, the nature of the effects typical of social science research often make it difficult to reject the null hypothesis, even when we are conducting studies with seemingly adequate sample sizes. The following examples from Hedges and Olkin (1985) make the point vividly.

Assume that we have a true effect in the population with a magnitude of $d = .50$, an effect Cohen (1977) defines as 'medium'. If we were to conduct repeated studies with samples of $N = 50$, fewer than 50 percent of these studies would be significant at the $p < .05$ level! The picture is even gloomier for smaller effects. If we have a population d of .20 (a 'small' effect according to Cohen), even with samples of $N = 100$, fewer than 20 percent of the studies would be statistically significant (Hedges & Olkin, 1985, p. 5). In these examples, a reliance on significance testing would lead us to conclude a majority of the times, *wrongly*, that an effect did not exist. A reliance on effect-size estimation would yield a more accurate picture of reality.

Effect Sizes, Replications and Meta-analysis

Chow's article contains a number of statements disparaging the utility of replication studies and meta-analysis. He claims that an 'overemphasis on replication studies, in fact, may be misleading when we are concerned with conceptual rigor' (p. 352), because successful replication studies may include the very same confounding factors that threaten internal validity. However, Chow's criticisms of 'mere' replications are based on a faulty and overly restrictive conception of what constitutes replication: 'By definition, the design and experimental task of the original study must be used in the replication studies' (p. 352). It is more accurate to speak of a continuum of replications, ranging from 'exact' replications using the same procedures to replications that employ radically different operationalizations (Rosenthal, 1990). The criticisms Chow raises are applicable only to exact replications, and exact replications are actually fairly rare in the behavioral sciences given the premium on journal space and subsequent unpublishability of such replications. Inexact replications help to provide the 'converging operations' that Chow acknowledges to be a crucial part of the theory-corroboration process.

Meta-analysis is a crucial tool for summarizing and integrating the outcomes of these replications. Chow is not a fan of meta-analysis, saying:

> Literature review is an intellectual exercise, the purpose of which is to examine whether or not a theory is properly supported by available evidence . . . . Consequently, advocating a standardized, numerical way of undertaking a literature review is antithetical to the purpose of reviewing research findings (p. 354).

I have always been struck by the illogic employed by critics of meta-analysis. A good meta-analysis does everything a traditional literature review does; it just does *more* by providing combined effect sizes and

significance levels in addition to a narrative analysis of the studies. A meta-analysis thus can be equally as much an intellectual exercise as a traditional literature review.

Chow and other critics worry that meta-analysis obscures truth by mixing together the results of dissimilar studies using radically different methods and variables. However, this criticism ignores the fact that a good meta-analysis will block by important theoretical and methodological variables. Through the judicious coding and contrast analyses of such variables, meta-analysis can help advance theory by testing hypotheses and identifying relationships that were previously undetected. Meta-analysis can also inform theoretical questions in a way that is impossible within single studies, because no single study can include as many manipulations or levels of an independent variable as can be assessed in a meta-analysis (Harris, 1991). Lastly, Cooper and Rosenthal (1980) have shown that meta-analysis can detect relationships in a body of literature that traditional literature reviews, with their emphasis on 'box scores', may miss.

Summary and Conclusions

Chow expresses well the heart of theory corroboration in the behavioral sciences when he says:

> . . . it is crucial for a researcher to ask (a) whether or not the set of data in question warrants the acceptance of the theory of interest, and (b) whether or not the data warrant the acceptance of any one of the competing theories (p. 348).

He claims that significance tests are all that is needed to answer these two questions. Chow's arguments are logically consistent. However, simply because the approach Chow advocates is logically permissible does not mean it is the best way of doing science.

Significance tests are not enough. They are useful for assuring us that our results are not due to random sampling fluctuation. Yet as scientists we should not stop there. Significance tests do not tell us anything about the nature or magnitude of a relationship, and the power levels typically employed in behavioral research mean that a reliance on significance tests will often lead to the incorrect conclusion that a relationship does not exist. The effect size associated with a focused contrast testing a hypothesis is the single best, most informative index of the relationship between theory and data. Comparison of effect sizes from contrasts predicted by competing theories is a better way than significance testing to judge the relative merits of the theories. Psychological theory is best advanced through estimating effect sizes, conducting replications to provide converging operationism

and performing meta-analyses on the replications so as to provide the most informative summary of the evidence bearing on the theory.

Notes

I would like to thank Rick Hoyle, Richard Milich and Michael Nietzel for their helpful comments, and my gratitude to Robert Rosenthal is enormous for his training in the philosophy toward data analysis described herein.

Requests for reprints should be addressed to Monica Harris, Department of Psychology, University of Kentucky, Lexington, KY 40506–0044, USA.

References

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

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin, 103*, 105–110.

Chow, S.L. (1989). Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin, 106*, 161–165.

Chow. S.L. (1991). Conceptual rigor versus practical impact. *Theory and Psychology, 1*(3), 337–360.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Cooper, H.M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*, 442–449.

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

Harris, M.J. (1991). Controversy and cumulation: Meta-analysis and research on interpersonal expectancy effects. *Personality and Social Psychology Bulletin, 17*, 316–322.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

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

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

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R. (1990). Replication in behavioral research. In J.W. Neuliep (Ed.), Handbook of replication research in the behavioral and social sciences, [Special Issue]. *Journal of Social Behavior and Personality, 5*, 1–30.

Rosenthal, R., & Rosnow, R.L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* Cambridge: Cambridge University Press.

MONICA HARRIS is an assistant professor of social psychology at the University of Kentucky, Lexington, KY 40506-0044, USA. Her two

primary research interests are interpersonal expectancy effects and meta-analysis. She is author of 'Controversy and cumulation: Meta-analysis and research on interpersonal expectancy effects', which appeared recently in the *Personality and Social Psychology Bulletin*.