Available online at www.sciencedirect.com
SCIENCE DIRECT®

Applied & Preventive Psychology 11 (2004) 69–72

APPLIED & PREVENTIVE PSYCHOLOGY

http://www.elsevier.com/locate/app

Commentary

Paul Meehl and the evolution of statistical methods in psychology

James H. Steiger

Department of Psychology and Human Development, #512 Peabody College, Vanderbilt University, Nashville, TN 37203, USA

Abstract

In his landmark 1978 paper, Paul Meehl delineated, with remarkable clarity, some fundamental challenges facing soft psychology as it attempts to test theory with data. In the quarter century that followed, Meehl's views stimulated much debate and progress, while continually evolving to keep pace with that progress. This paper pays homage to Meehl's prescience, and traces the impact of his ideas on the recent shift of emphasis away from hypothesis testing and toward confidence interval estimates of effect size.

Keywords: Paul Meehl; Statistical methods; Psychology

1. Introduction

Paul Meehl's (1978) paper on statistical testing displays, in rich abundance, the remarkable characteristics that made him one of the giants of modern psychology. First, there is the astonishing depth and breadth of his scholarship. Meehl read enough for any 10 of us, and he read deeply. Second, there is the powerful sense of purpose, manifested both in the sparks of his wit, and the way he constantly drove toward solutions while acknowledging the complexity of problems. Finally, there is his enormous *perspective*, his "ability to sort out what is important and what is right" (Steiger & Fouladi, 1997, p. 221).

Meehl (1978) delivered a devastating critique of the way many soft psychologists employed statistical testing in theory validation. Several key ideas emerged:

1. Nil hypotheses[1] (hypotheses of absolutely no mean difference or precisely zero correlation) are always false in soft psychology, so the percentage of rejections is largely a function of statistical power.
2. When we test a theory, we are usually testing something more—the theory plus several ancillaries (which Meehl symbolized as (T, A, C)). Falsifying the conjunction of the theory and its ancillaries need not be fatal for the theory—it may simply indicate that the ancillaries need fixing.
3. Finding statistical evidence that agrees with a theory does not prove the theory is true—this is the fallacy of *affirming the consequent*.
4. Consequently, when evaluating a theory with hypothesis tests, we must give much greater weight to negative results than to positive ones.
5. Running a group of tests and "counting heads" (i.e., polling agreements and disagreements in many hypothesis tests to validate a theory) ignores principles (1)–(4), and should be discontinued.
6. In spite of (3), Meehl felt that when (T, A, C) generates a "high risk numerical point prediction" and data agree with the prediction, then it should strike us as implausible that the theory is wrong.

In what follows, I will reflect briefly on some of the key ideas in Meehl's work, how they have impacted on my own thinking, and how they helped stimulate substantial progress in the way we test statistical hypotheses.

2. Noncentrality interval estimation and tests of close fit

Meehl's (1978) paper concentrated on improper approaches to "Reject-Support" testing, in which rejecting a null hypothesis supported an experimenter's theory. However, many important model-evaluation efforts at the time were attempting to employ "Accept-Support" logic, in which

E-mail address: james.h.steiger@vanderbilt.edu (J.H. Steiger).

[1] Meehl used the term "null hypothesis," but he generally was referring to a hypothesis of a nil parameter value, i.e., zero mean difference or zero correlation. Modern statistical testing assumes a broader meaning for "null hypothesis," so for clarity I use the more exact term "nil hypothesis" when warranted.

0962-1849/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.appsy.2004.02.012

accepting a statistical null (or nil) hypothesis supports the researcher's model. The problems Meehl identified in the context of Reject-Support testing were, if anything, more severe in Accept-Support testing. For example, in 1978 structural equation modeling was becoming popular in psychology. Statistical testing in structural modeling involved an "Accept-Support" rationale: a model was supported if a chi-square test of perfect fit was not rejected. There were many problems with this rationale. In particular, if a model had near perfect fit, it would still be rejected if sample size were large enough. So in essence, high precision worked against the researcher trying to verify a model.

I read Meehl's (1978) paper just after it appeared, and was profoundly influenced by its emphasis on the interplay between logic and statistics. It caused me to reflect on some of the fallacies of nil hypothesis testing in structural modeling and factor analysis. A partial solution to the dilemma of near-perfect fit suggested itself. Instead of testing whether fit is perfect, why not assess how good (or bad) it is with a confidence interval? I created an index of badness of model fit (the RMSEA) that combined model fit and model complexity, and showed how it can be estimated with a confidence interval. If an entire RMSEA confidence interval (Steiger & Lind, 1980) fell within reasonable distance of perfect fit, then fit could be assumed to be good-enough for practical purposes, even though the nil hypothesis was rejected at the 001 level. As sample size (and precision) increase, the confidence interval gets narrower, and a worthy (very good but not perfect) structural model is more likely to be declared reasonable.

Similar notions were emerging in diverse areas of statistical testing around that time. For example, Westlake (1976) had already proposed a confidence interval based approach to bioequivalence testing, in which one attempts to decide whether levels of a drug are within established limits. Westlake suggested examining a single confidence interval, and seeing whether the entire interval fell within the bioequivalence limits. Bioequivalence need not be perfect, as long as the confidence interval showed it was good-enough. A few years later, Fleishman (1980) discussed confidence interval estimation of various effect size indices in connection with one-way fixed-effects ANOVA.

In 1985, Serlin and Lapsley proposed the "good-enough principle," which recommended formal tests of close fit (or not-close fit) to replace tests of perfect fit. Serlin and Lapsley pointed out that Meehl had not considered the "good-enough" principle in his 1978 critique, and that this principle resolved many of the problems Meehl had discussed. Meehl (1990) responded to this criticism by frankly admitting that Serlin and Lapsley had a point, but insisting further that key problems remained.

Effect-size interval estimation and tests of close fit were slow to gain adherents in psychology in the areas where they were most needed-tests on means and regression analysis. On the other hand, the same ideas became quite popular in the context of structural equation modeling. The RMSEA and related noncentrality-based measures of fit (Browne & Cudeck, 1993; Steiger, 1989, 1990a) were adopted in the 1990s in all the commercial structural modeling software programs. Soon many psychologists found themselves (at least loosely) employing the Serlin & Lapsley (1985) good-enough principle while perhaps not explicitly recognizing it.

I remained convinced that confidence interval estimation had broader possibilities, and tried to spread the word in a series of conference presentations (Steiger, 1990b, 1990c). Encouraged by conversations with a number of colleagues at these conferences (especially Jack Cohen and Paul Horst), Rachel Fouladi and I completed work on R2 (Steiger & Fouladi, 1992), a program that performed nonstandard hypothesis tests and exact confidence interval estimates on the squared multiple correlation. Around the same time, Serlin and Lapsley (1993) described procedures for implementing the good-enough principle with tests on means, and described in general terms how the good-enough principle could be adapted for a variety of tests, including tests of close fit and not-close fit. Unfortunately, they did not provide functional software, and their excellent recommendations did not receive as much attention as they deserved. Cohen (1994), in an extremely influential paper in the *American Psychologist*, urged a move away from hypothesis tests and toward confidence interval estimates of effect size, and Schmidt and Hunter (1997) (see also Schmidt (1996)) delivered strong attacks on hypothesis testing. Soon, the APA convened a special panel (on which Paul Meehl served as a Senior Advisor) to examine significance testing and make recommendations.

While the APA panel was deliberating, calls for a shift toward interval estimation and tests of close fit increased. MacCallum, Browne, and Sugawara (1996) discussed power calculation and tests of close fit and not-close fit in structural equation modeling, using my RMSEA index. Steiger and Fouladi (1997) described *noncentrality interval estimation*, a general approach to effect size confidence intervals that could replace hypothesis tests in a wide variety of common statistical tests (especially ANOVA and multiple regression). Steiger (1999) presented a general computer program that implements these interval estimation procedures.

Interestingly, Meehl's 1997 chapter in the same book (Harlow, Mulaik, & Steiger, 1997) as Steiger and Fouladi (1997) also urged a shift of emphasis toward the use of confidence intervals. The final report of the APA Task Force on Statistical Inference (Wilkinson, 1999) recommended that confidence intervals always be reported in connection with effect size estimates and measures of correlation. In 2001, *Educational and Psychological Measurement* published a special issue with a series of tutorial papers (e.g., Cummings & Finch, 2001; Smithson, 2001) on noncentrality-based interval estimation procedures.

In a paper (Steiger, 2004) that is about to appear in *Psychological Methods*, I show how interval estimation and "good-enough" statistical testing can be united to produce a rational approach to evaluating omnibus effects and focused contrasts in fixed-effects ANOVA. These techniques provide all the information available in a traditional hypothesis test, and more. The paper also discusses extensions to multivariate analysis, random effects models, and regression modeling.

3. Barriers to progress

The preceding section describes the progress that has been made in developing procedures for quantifying badness of fit of a statistical model, and how many of the early developments in this field were stimulated by Meehl's (1978) critique. We now have a much better idea what we *should* do in regression and ANOVA. What the majority of researchers *will* do it is, unfortunately, determined to a considerable extent by what SPSS, SAS, and other manufacturers of general-purpose statistical software are willing to implement in their programs. For better or worse, statistical practice is software-driven (Steiger, 2001).

To accompany the publication of Steiger (2004), my colleague Rachel Fouladi and I will be distributing Windows freeware to perform the calculations necessary to implement noncentrality interval estimation techniques, hopefully thereby eliminating some of the barriers to their use. Ultimately, however, journal editors and authors need to respond to the availability of such tools.

4. Future challenges

In the decades that followed his 1978 paper, Meehl (1990, 1997) refined and sharpened his views considerably. This evolution occurred partly in response to technical developments that Meehl's own work had helped stimulate. Meehl's 1997 review emphasized confidence interval estimation as a superior alternative to hypothesis testing, and attempted, via a "Corroborative Index," to quantify and formalize his views on how theories should be evaluated. The Corroborative Index attempted to combine sophisticated notions of "closeness" of the data to prediction, and the ability of a theory to tolerate deviations from prediction. As Meehl (1997) pointed out, a shift toward interval estimates of effect size only partly resolves the dilemmas facing psychologists as they attempt to validate theories. A key issue is how genuinely risky the original prediction was—high accuracy in a low risk prediction does not tell us much.

Assessing riskiness of a model's predictions and the corroborative value of close fit is complicated by some perplexing statistical problems: (a) the generalized ability of a particular model to fit many data sets well, and (b) the existence of equivalent models, the fact that several different models may fit a data set equally well. Statisticians have just begun to formalize and quantify these problems. Meehl and Waller (2002) described an extraordinarily creative (if controversial) attempt to reconcile these problems in the assessment of path models. To see Meehl, in his 80s and in ill health, collaborating so vigorously and productively with a gifted colleague only slightly more than half his age, should be an inspiration to all of us to carry his work forward, and reaffirm his commitment to a logical, quantitatively based science as we train the next generation of psychologists.

References

Browne, M. W, & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cummings, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

Fleishman, A. E. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40,* 659–670.

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

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

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.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

Meehl, P. E., & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods, 7,* 283–300.

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

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? Mahwah, NJ: Lawrence Erlbaum Associates.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological theory: The good-enough principle. *American Psychologist, 40,* 73–82.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences*: *Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 603–630.

Steiger, J. H. (1989). *EzPATH*: *A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT Inc.

Steiger, J. H. (1990a). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Steiger, J. H. (1990b, October). *Noncentrality interval estimation and the evaluation of statistical models*. Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology, Newport, RI.

Steiger, J. H. (1990c, June). *Noncentrality interval estimation in the analysis of covariance structures*. Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96,* 331–338.

Steiger, J. H. (2004). Beyond the F-test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, in press.

Steiger, J. H., & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers, 4*, 581–582.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In: L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 251–257). Mahwah, NJ: Lawrence Erlbaum Associates.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the May annual meeting of the Psychometric Society, Iowa City, IA.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32*, 741–744.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and Explanations. *American Psychologist, 54*, 594–604.