Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied & Preventive Psychology 11 (2004) 61–64
APPLIED & PREVENTIVE PSYCHOLOGY
http://www.elsevier.com/locate/app

Commentary

Another quasi-30 years of slow progress

Gregory A. Miller

Departments of Psychology and Psychiatry, Beckman Institute Biomedical Imaging Center, University of Illinois at Urbana-Champaign, 2100 S. Goodwin, Urbana, IL 61801, USA

Abstract

Meehl (1978) discussed a variety of characteristics of the culture of scholarly psychology that critically affect its progress toward a stronger science. These characteristics include assumptions about the nature of theory, approaches to the testing of theories, and the reliance on significance tests that is pervasive in many subfields of psychology research. Several of these characteristics and Meehl's perspective on them are examined years later in this brief commentary. Meehl's criticisms, though sometimes misrepresented, remain compelling and strikingly current. Yet it should be remembered that Meehl emphasized that "soft" psychology at its best is a profound and worthy challenge and will necessarily progress slowly. It can be improved, but not replaced, by hardnosed scholarship.

Keywords: Meehl; Soft psychology; Significance testing; ANCOVA; Power analysis; Randomization; Psychodynamic

What a joy to read this paper (Meehl, 1978) on the slow progress of soft psychology again. The first paragraph ends with the bait that, if the article gets us to think about the relationship between theory-testing and significance-testing, "this article will have served its scholarly function" (p. 806). It certainly did get us thinking. Beyond the provocative content, over and over in reading it I recognized certain wonderfully tailored phrases that had stayed with me from earlier readings. It would be tempting to devote the present very short commentary to praising favorite gems among the content and phrasing, but instead this note examines a few items from, and adds to, Meehl's list of factors keeping the progress slow.[1]

(1) Meehl (1978) cited 20 reasons that render "soft" psychology a truly difficult science. Under "nuisance variables", he referred to difficulty in either statistically or causally separating variables that are so intertwined with phenomena of interest that we cannot distinguish them (see also Meehl, 1970, 1971). We can highlight Meehl's concerns in two directions, re: the construct validity of residuals and the possible role of a covariate in substantive theory. Routinely one sees authors of publications and grant applications referring to "statistically controlling" for or "removing the effect of" some covariate, without carefully considering the consequences. If the variables of interest are meaningfully related to the covariate, then potentially one has removed not just nuisance variance but meaningful variance—one no longer has the variables of interest available for analysis. An example of this problem is the common, inappropriate use of ANCOVA when groups differ on the covariate. Removing such variance does not "control for" or "correct for" anything. It just removes variance. Whether that removal has done violence to the variables of interest—whether the residualized variables are still construct-valid—has to be carefully considered (see Miller & Chapman, 2001, for extended discussion). This problem at first glance appears to be an enormous inconvenience for experimental psychopathologists, whose subject groups so often differ in ways not seen as central to the question at hand. Nevertheless, what is needed in the face of such inconvenience is not arithmetic but conceptual re-evaluation of the original variables. Possibly what appears to be a nuisance variable actually warrants a nontrivial role in one's theory.

(2) Meehl (1978, p. 806) denounced some theorizing as weak: "... most so-called 'theories' in the soft areas of psychology ... are scientifically unimpressive

E-mail address: gamiller@uiuc.edu (G.A. Miller).

[1] Although Meehl (1978) relied on the common "hard science"/"soft science" distinction in his choice of terms, he made a strong case for "soft psychology" actually being the harder side of the field—asking the more difficult questions.

0962-1849/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.appsy.2004.02.010

and technologically worthless". We can go further: most "models" or "theories" we run across are not theories or models, they are lists of concepts (sometimes very appropriate ones) with too little spelled-out mechanism sewing them together in some credibly dynamic way to be called a theory or model. In the psychopathology literature (e.g., Miller, 1995), the pervasive "diathesis-stress model" is not a model. It is a proposal that two factors arise in temporal sequence with some combined effect on subsequent dysfunction. In most cases, it carries the implication that different factors have different time courses, with the diathesis relatively static and the stressor acute or cumulative. It is a fine proposal, surely true in many cases. But without also spelling out the mechanisms in the relationship between the diathesis, the stressor, and the dysfunction, we do not have a model; no model, no test of the model; no test, no danger of refutation. Meehl would not be pleased, because he championed Popperian risk of refutation as key to scientific progress.

(3) A message that Meehl (1978) particularly hammered home is that significance testing is often the wrong thing to do entirely. More subtly, it can distract us from the experiment itself. After 6 days punching a calculator computing *t*-tests for my undergraduate honors thesis, I brought the results to my advisor, with much pride and excitement, because a few of them were actually "significant" at the 0.05 level. He immediately asked one question I had not anticipated. "What are the effects?", by which he meant: what are the means contributing to those *t*-tests, and what is the direction of their differences? I did not know; I had not noticed. He was a bit exasperated. It had not occurred to me that, to make any sense at all of the inferential statistics, one needs to start with the descriptive statistics. Our first task is to describe what we see in our experiment—not to estimate whether, if someone else somewhere else had done a different experiment, they would have seen the same thing. Perhaps my most common recommendation as a journal editor and reviewer is a thorough rewrite of a Results section so that paragraphs generally lead with the important descriptive findings, rather than the inferential statistics. Many of us were taught that the descriptive statistics, if not "significant", must be banished from consideration. But this precludes discovery of fortuitous results. One plans an experiment with due diligence, but once done one must interrogate the data to discover what experiment actually occurred—not just what the results were but what the manipulation was. I usually measure something biological in my research, obtained during some behavioral task. Before looking at the psychophysiological scores (fMRI, EEG, autonomic measures), I want to start with the behavioral performance data, to see what they tell me about what my subjects really did. If such a manipulation check suggests that my subjects underwent a procedure different from my intention, I do not throw out the data, I investigate whether what they did was largely consistent across subjects and largely interpretable. If so, I have a valuable data set after all. If not internally consistent and interpretable, little else matters. All of that trumps significance levels.

(4) Meehl (1978, p. 807) lamented "... a disturbing absence of the *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics." Surely a major reason for the lack of accumulation is that we (still) do not have nearly the consensus on the set of phenomena our field should focus on that some other disciplines have. The making of, and consequences of finding, new, superheavy elements has been a central focus of chemistry and physics for 50 years. Success may lead to stable, super-dense materials of enormous practical significance. What comparable goal does soft psychology have? "Curing mental illness" sounds wonderful, but we are far from agreed on what constitutes mental illness and even whether "cure" is the appropriate metaphor. (What should successful psychotherapy look like? What does a good life look like?—Not questions that inferential statistics will answer, despite very good progress on empirically supported psychotherapy in recent decades.) We should not be fooled by logically impossible claims, for example, that schizophrenia is a brain disease (c.f. NIMH web site) or by a putative biological theory of schizophrenia that is not a theory of schizophrenia at all. The dopamine theory of schizophrenia was an impressive account of dopamine phenomena associated with schizophrenia, lacking any adequate mechanistic account of how dopamine dysfunction could account for psychological symptoms. And it is psychological symptoms that define schizophrenia, not only operationally but essentially (see Miller, 1996, for extended discussion).

(5) Precious little attention was paid to statistical power in 1978. Grant applications now routinely address it, but most get it wrong. The modal discussion is a perfunctory treatment of a tractable subset of the hypotheses, with the power to confirm them based on some pre-existing (even arbitrary, such as "medium"; Cohen, 1988) estimate of the likely effect size. This is not a relevant power analysis. If we respect and extend Meehl's (1978, p. 814, 822) pioneering discussion of power, what is needed is a clear stand on how big an effect is *worth* finding, not how big an effect is *likely*—we need to know what the power is for an *adequate* test of the hypothesis. The typical hypothesis is that two means differ (or more generally that two samples differ in *some* way), whereas Meehl repeatedly pointed out that the far more valuable prediction would be about the magnitude of the difference. So we have progressed since 1978 in that we now often invoke statistical power, but it is not clear that we understand it any better than when Meehl drew our attention to it.

(6) Meehl (1978) provided a self-reflective appendix on his faith in aspects of psychoanalysis, at a time when psychodynamic and behavioral approaches were fighting for the soul of academic clinical psychology. He stressed the importance of distinguishing whether a theory is testable now with available methods versus potentially testable with future methods, the latter being sufficient for good science. He noted that behavior therapy had proven both effective and insufficient. Years later, the academic fight is over. Perhaps ironically, like the progeny of the French Norman knights who, having conquered England, came to speak English, behavior therapy conquered academic clinical psychology, with the result that now essentially everyone is psychodynamic: the classical definition of "psychodynamic" does not entail a Freudian architecture of id, etc. only the assumption that there are "forces in the mind" (Greenson, 1967, p. 23). Any self-described behaviorist who uses concepts of cognition, emotion, or motivation that have any meaning beyond observables (thus, essentially all modern behaviorally oriented clinical psychologists; see Kozak & Miller, 1982; Miller & Kozak, 1993) clearly relies on psychodynamic concepts. If one broadens "psychological" beyond the intrapersonal to include interpersonal process (What clinician does not attend to interpersonal dynamics?), then the case is even easier to make, whether or not one calls it "transference". Meehl's faith in aspects (not all) of psychodynamic heritage is no less defensible than in 1978—if anything, the subsequent evolution of the field has validated that faith.

(7) With all due respect to Meehl (and distinguishing him from the slash-and-burn attitude that has been misread into the 1978 paper), inferential statistics are not the final common pathway to evil or bad science. When I was in graduate school, essentially no one in psychology did MANOVAs. Only one unwieldy program was (not widely) available to do it. Van Egeren (1973) had radically proposed reliance on MANOVA in psychophysiology research, to no avail. Vasey and Thayer (1987) trumpeted its value half a generation later, when it was finally widely available, but still one frequently encounters the misunderstanding that its relevance is only when there are multiple dependent variables to be crunched together. Now 31 years after Van Egeren, MANOVA is still relatively rare. As MANOVA has been so long in coming, I suspect that the current sleeper in inferential statistics is randomization, bootstrapping, permutations, and similar methods (e.g., Maris, 2004; Wasserman & Bockenholt, 1989). These have the beauty of allowing the investigator to design a boutique statistic, customized for the hypothesis at hand. For example, rather than doing a giant omnibus ANOVA (or MANOVA!), then a series of simple-effects analyses that systematically throw away information, one defines a specific relationship in the data, computes its observed magnitude, and estimates the probability of such a magnitude arising by chance. Meehl might not fully approve; but surely he would condone a strategy that forces one to think very carefully and explicitly about testing one's hypothesis. The omnibus ANOVA/MANOVA with many cells in the design is like throwing a deck of cards into the air and then seeing what hand one can play with those that turn face-up. Bootstrap techniques are not yet popular, I suspect, exactly because of the burden they place on the investigator, to specify (and to code in software) a particular relationship among variables. And the software to do bootstrapping is not yet built into SPSS (though it is in SAS). But the flexibility of this strategy will surely prevail eventually. Meehl might even approve, because one's boutique statistic can represent quantitatively specific relationships far more precise than "from the same population".

(8) It might not be a stretch to read Meehl (1978) as hoping that soft psychology will someday progress past any reliance on significance testing. As a new assistant professor I sometimes rode the bus to campus with a famous senior colleague, biopsychologist Phil Teitelbaum. One day he explained why he had no use for inferential statistics. In his lab, he tinkered with a paradigm to explore a phenomenon until he understood it. Once he understood it, the phenomenon would have such a whopping effect size, such good consistency across subjects (brain-lesioned rodents), that no statistical inference would be necessary. He believed that, if one needs inferential statistics to draw a conclusion, the conclusion is premature. This echoes Meehl (1978, p. 825) noting that most physicists have no use for statistical significance tests, they have quantitative models, precise instruments, and enough observations to judge, by simple inspection, whether the data compellingly fit the model. We will have to have consensus on adequate models, as well as construct-valid measures, if we are to achieve that. Yet statistical inference is no substitute for adequate models and good measures. Meehl argued that some of the problems that "soft" psychology tackles are remarkably complex and thus difficult to capture with straightforward models and measures.

(9) If the harder sides of psychology are closer to the harder sciences, one might wonder whether the growing presence in psychology of genetics and neuroscience will lead the field to greater rigor—or will suffocate those parts of the field less amenable to such rigor—and whether the latter is a good thing. Meehl (1978) pleaded that soft psychology not turn its back on the hard questions it asks, in favor of easier questions. Meehl (1978) is often cited in support of casual derision of softer psychology, but the paper takes great pains to prevent that: "... it is usually not possible in the soft areas of social science to provide rigorous, explicit, or ... operational definitions for theoretical concepts" (p. 815). What a radical premise! With Meehl, let us grant that premise

and, without embarrassment, accept that the progress will be slow.

Acknowledgements

The writing of this manuscript was supported by grants from the National Institutes of Health (R01 MH61358, R01 MH65429, R21 DA14111, T32 MH14257, T32 MH19554) and the University of Illinois Intercampus Research Initiative in Biotechnology. Helpful comments on an earlier draft were received from Bruce N. Cuthbert, J. Christopher Edgar, Wendy Heller, Michael J. Kozak, and David A. Smith.

References

Cohen, J. (1988). *Statistical power analysis for the social sciences.* Hillsdale, NJ: Erlbaum.

Greenson, R. R. (1967). *The technique and practice of psychoanalysis (Vol. 1)*. New York: International Universities Press.

Kozak, M. J., & Miller, G. A. (1982). Hypothetical constructs versus intervening variables: A re-appraisal of the three-systems model of anxiety assessment. *Behavioral Assessment, 14,* 347–358.

Maris, E. (2004). Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology, 41,* 142–151.

Meehl, P. E. (1970). Nuisance variables and the ex post factor design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: Analyses of theories and methods of physics and psychology* (*Vol. 4,* pp. 373–402). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1971). High school yearbooks: A reply to Schwartz. *Journal of Abnormal Psychology, 77,* 143–148.

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.

Miller, G. A. (Ed.) (1995). *The behavioral high-risk paradigm in psychopathology.* New York: Springer.

Miller, G. A. (1996). Presidential address: How we think about cognition, emotion, and biology in psychopathology. *Psychophysiology, 33,* 615–628.

Miller, G. A., & Chapman, J. P. (2001). Invited paper: Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48.

Miller, G. A., & Kozak, M. J. (1993). A philosophy for the study of emotion: Three-systems theory. In: N. Birbaumer & A. Öhman (Eds.), *The structure of emotion: Physiological, cognitive and clinical aspects* (pp. 31–47). Seattle: Hogrefe and Huber.

Van Egeren, L. F. (1973). Multivariate statistical analysis. *Psychophysiology, 10,* 517–532.

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measaures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology, 24,* 479–486.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology, 26,* 208–221.