Available online at www.sciencedirect.com

SCIENCE DIRECT®

Applied & Preventive Psychology 11 (2004) 83–86

APPLIED & PREVENTIVE PSYCHOLOGY

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

Commentary

The fallacy of the null hypothesis in soft psychology

Niels G. Waller

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

Abstract

In his classic article on the fallacy of the null hypothesis in soft psychology [J. Consult. Clin. Psychol. 46 (1978)], Paul Meehl claimed that, in nonexperimental settings, the probability of rejecting the null hypothesis of nil group differences in favor of a directional alternative was 0.50—a value that is an order of magnitude higher than the customary Type I error rate. In a series of real data simulations, using Minnesota Multiphasic Personality Inventory-Revised (MMPI-2) data collected from more than 80,000 individuals, I found strong support for Meehl's claim.

Keywords: Null hypothesis; Soft psychology; MMPI-2

Meehl's (1978) classic article on the fallacy of the null hypothesis in soft psychology was neither his first (Meehl, 1967) nor his most comprehensive (Meehl, 1990a, 1997) treatment of the proper role of significance testing in theory appraisal. Nevertheless, this is Meehl in top form, and the article remains one of my favorites. My fondness for the "two knights" stems, in part, from personal history: this was my first Paul Meehl article. It was also my first encounter with Meehl's awe inspiring brilliance.[1] That initial jousting match more than 20 years ago taught me many important concepts, such as 'construct validity', 'verisimilitude', 'consistency tests', and 'taxometrics.' At the time I was unaware of the role that these concepts—and the man who lobbied so vociferously for their adoption—would play in my subsequent professional development (cf. Meehl & Waller, 2002; Waller & Meehl, 1998, 2002).

Page constraints force me to limit my comments to only one of the many interesting points that are contained in Meehl's article. The topic I have chosen is Meehl's belief that, within the soft areas of psychology—i.e., those areas in which random assignment to treatment and control groups is infeasible due to practical and/or ethical constraints—the null hypothesis of nil group differences is almost always false. That being the case, according to Meehl, null hypothesis refutation provides feeble support for theory confirmation. Of course, Meehl expressed this idea more colorfully.

> . . . Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe that 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. (Meehl, 1978, p. 817)

Meehl was not the first to raise these concerns (cf. Hogben, 1957; Rozeboom, 1960; see Bill Thompson's web site for a compilation of over 400 references on the topic: http://www.biology.uark.edu/coop/Courses/thompson5.html). Nevertheless, his ability to couch this debate within the larger framework of philosophy of science and theory appraisal was the needed impetus to get scores of research psychologists thinking about this unsettling issue (Meehl's paper is a citation classic that has been reprinted in numerous outlets).

Twenty-five years ago the notion that "the null hypothesis, taken literally, is always false" (Meehl, 1978, p. 822; see similar statements in Meehl, 1967, 1990a, 1990b, 1997) was a controversial claim. Today, the idea is less iconoclastic (Cohen, 1994). Looking back it is interesting to consider why Meehl was so troubled by this realization. Reflecting on the slow rate of progress in clinical psychology and related disciplines, Meehl realized that if the null hypothesis

E-mail address: niels.waller@vanderbilt.edu (N.G. Waller).

[1] I first learned of Meehl's article in a methods course taught by Brendan Maher. Several years earlier Maher had edited the special issue of the *Journal of Consulting and Clinical Psychology* in which Meehl (1978) first appeared.

0962-1849/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.appsy.2004.02.015

Table 1
Descriptive statistics for sample data

	N	Age				Assessment setting (%)						
		Minimum	Maximum	Mean (S.D.)	Median	1	2	3	4	5	6	7
Males	41,491	16	95	37.22 (11.46)	36.00	55	05	06	05	04	01	24
Females	39,994	16	91	36.92 (11.42)	30.00	56	07	06	06	01	01	22

1: outpatient mental health; 2: inpatient mental health; 3: general medical; 4: chronic pain; 5: correctional; 6: college counseling; 7: other.

of nil group differences is always false then (given sufficient power) directional hypothesis tests in these disciplines have a probability of one half of reaching statistical significance. In other words, the a priori probability of rejecting the null hypothesis in soft psychology is 10 times higher than the presumed Type I error rate. In nonexperimental settings, this implies that null hypothesis significance testing (NHST) is a feeble method for corroborating substantive theories because it fails to subject theories to grave danger of refutation (Popper, 1959).

Some researchers were troubled by this conclusion (Keuth, 1973, Oakes, 1975; though see Bakan, 1966). Meehl's views, however, were not merely speculations from the armchair. In his first contribution to the significance testing controversy (Meehl, 1967), Meehl noted that when he and Lykken analyzed data from over 55,000 Minnesota high school seniors they found "statistically significant relationships in 91% of pairwise associations among congeries of 45 miscellaneous variables such as sex, birth order, religious preferences, number of siblings, vocation choice, club membership, college choice, mother's education, dancing, interests in wood working, liking for school, and the like" (p. 109). In a later informal study using Minnesota Multiphasic Personality Inventory (MMPI) data, Meehl (1997) reported that "[w]ith a sufficiently large sample, almost all of the 550 items of the MMPI are significant correlates of gender" (p. 405). These and other results convinced Meehl that "in social science everything correlates with almost everything else, theory aside" (Meehl, 1997, p. 393). David Lykken (1991) coined a name for this phenomenon: *the crud factor*.

During my stint in Minnesota, I was taught by Paul Meehl and other notable teachers to question sweeping claims until those claims were empirically corroborated through systematic investigation. Thus, in the spirit of this wise counsel, I conducted several experiments (real data simulations) to test Meehl's conjecture on the fallacy of the null hypothesis and its unfortunate consequence for theory testing.

Before running the experiments I realized that, to be fair to Meehl, I needed a large data set with a broad range of biosocial variables. Fortunately, I had access to data from 81,485 individuals who earlier had completed the 567 items of the Minnesota Multiphasic Personality Inventory-Revised (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). The MMPI-2, in my opinion, is an ideal vehicle for testing Meehl's claim because it includes items in such varied content domains as general health concerns; personal habits and interests; attitudes towards sex, marriage, and family; affective functioning; normal range personality; and extreme manifestations of psychopathology (for a more complete description of the latent content of the MMPI, see Waller, 1999). The data were obtained from the University of Minnesota Press.[2] For the studies reported below, an item response vector was deemed valid and included in the final analyses if it passed purposively conservative validity checks. The original data set included MMPI-2 protocols from 98,282 individuals. To be included in the final analyses, each record had to meet the following criteria: cannot say score ≤ 15, Lie T-score < 80, TRIN < 13, TRIN > 5, VRIN < 13, and back F scale < 90. Table 1 reports summary statistics for the final sample.

To set the stage for the first experiment, imagine that you are a gender theorist who is interested in sex differences. Further imagine that you have concocted a theory that is sufficiently powerful to generate a directional (alternative) hypothesis, such that the girls will score higher than the boys on a variable deemed important to your work.

Given this sketchy outline—and it is sketchy, notice that we have yet to specify the dependent variable—can we predict the probability of rejecting the null hypothesis of no group differences? Meehl suggests that we can. According to Meehl, with sufficient power, the probability of rejecting the null in favor of a directional alternative hypothesis is one half *irrespective of the dependent variable*. To test this claim, I programmed a computer to run the aforementioned hypothetical experiment 511 times on the MMPI-2 data.[3]

The computer program was structured as follows. First, a virtual coin was tossed to determine the direction of the alternative hypothesis. Next, the computer selected (without replacement) a random item from the pool of MMPI-2 items. Using data from the 41,491 males and 39,994 females, it then (a) performed a difference of proportions test on the item group means; (b) recorded the signed z-value; and (c) recorded the associated significance level. Finally, the program tallied the number of "significant" test results (i.e., those with $|z| \geq 1.96$). The results of this mini simulation were enlightening and in excellent accord with the outcome of Meehl's gedanken experiment. Specifically, 46% of the

[2] The author thanks Beverly Kaemmer, of the University of Minnesota Press, for making these data available.

[3] The 56 items of the Mf scale were not studied in this experiment; they were originally included in the MMPI because of their ability to discriminate between the sexes.

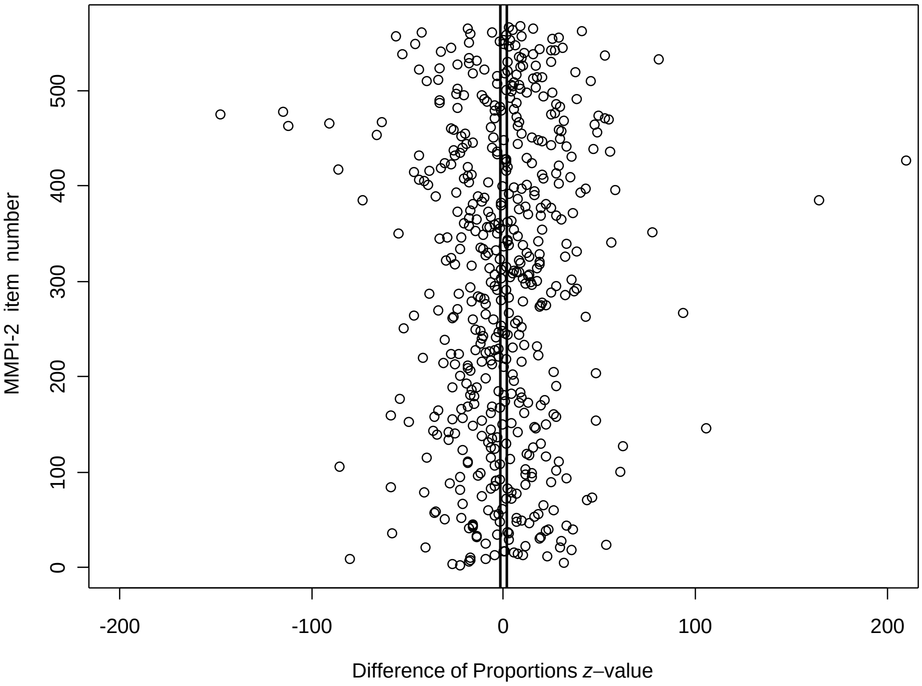

Fig. 1. Distribution of *z*-values for 511 hypothesis tests.

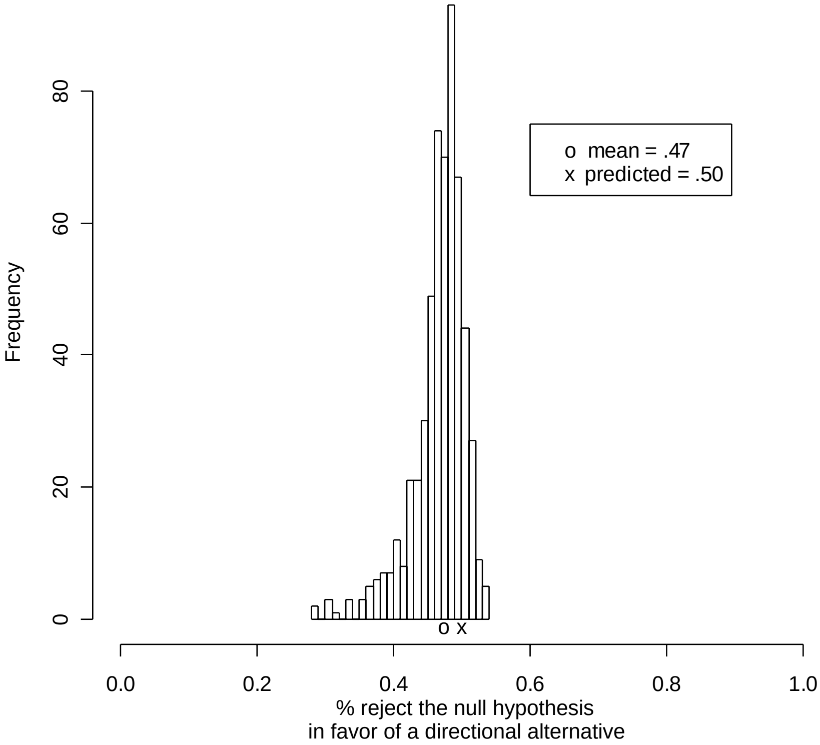

Fig. 2. Distribution of the frequency of rejected null hypotheses, in favor of a randomly chosen directional alternative, in 320,922 hypothesis tests.

directional hypotheses were supported at significance levels that far exceeded traditional *p*-value cutoffs. A summary of the results is portrayed in Fig. 1. Notice in this figure, which displays the distribution of *z*-values for the 511 tests, that many of the item mean differences were 50–100 times larger than their associated standard errors!

Admittedly, these findings replicate rather than extend Meehl's informal findings using the MMPI. Thus, we have yet to test the generality of Meehl's claim. We can easily broaden the scope of our test, however, by repeating the aforementioned study using samples formed by other grouping variables. The most obvious grouping variables within easy reach are the actual MMPI-2 items. In other words, we can assign individuals to groups on the basis of whether they answered TRUE or FALSE to a given item. By following this simple plan we can conduct an additional 567 simulation studies (one study for each item). Moreover, within each study, we can use the remaining items to conduct 566 hypothesis tests of no group differences (with a directional alternative determined by a virtual coin toss).

To simplify the programming task, I carried out the experiments just described using data from only the female subjects. Although this decision reduced my overall sample size by 50%, the remaining 39,994 cases should be more than sufficient to test the generality of Meehl's claim. Fig. 2 displays a summary of the findings from the 567 simulations. The histogram reports the percentage of rejected null hypotheses for 320,922 tests (567 items × 566 tests) that were run in this phase of the study. The circle at the base of the histogram marks the mean rejection rate (median $= 0.47$, mean $= 0.47$) whereas the "×" marks Meehl's predicted value of 0.50.

These findings provide strong support for Meehl's conjecture that in the soft areas of psychology the a priori probability of rejecting the null hypothesis of nil group differences is 0.50. My hope is that they will add another voice to Meehl's call for more rigorous methods for subjecting psychological theories to risky tests. Fortunately, researchers who wish to heed this call do not need to search for novel methods in the dark. Meehl (1990a, 1997, see also Meehl & Waller, 2002) has already illuminated the path we should take if we wish to increase the rate of progress in soft psychology.

References

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

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota multiphasic personality inventory*-2 (*MMPI-2*): *Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Cohen, H. (1994). The earth is round ($p < 0.05$). *The American Psychologist, 49*, 997–1003.

Hogben, L. (1957). *Statistical theory*. London, UK: Allen and Unwin.

Keuth, H. (1973). On prior probabilities of rejecting statistical hypotheses. *Philosophy of Science, 40*, 538–546.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology* (vol. 1). Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

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

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry, 1*, 108–141, 173–180.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244 [also in R. E. Snow & D. Wiley (Eds.) (1991). *Improving inquiry in social science*: *A volume in honor of Lee J. Cronbach* (pp. 13–59). Hillsdale, NJ: Lawrence Erlbaum Associates].

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests*? (pp. 393–425). Mahwah, NJ: Erlbaum.

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.

Oakes, W. F. (1975). On the alleged falsity of the null hypothesis. *Psychological Record, 25*, 265–272.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

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

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures*: *Distinguishing types from continua*. Newbury Park, CA: Sage.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. Embretson & S. Hershberger (Eds.), The new rules of measurement: What every psychologist and educator should know (pp. 185–217). Mahwah, New Jersey: Lawrence Erlbaum Associates, INC.

Waller, N. G., & Meehl, P. E. (2002). Risky tests, verisimilitude, and path analysis. *Psychological Methods, 7*, 323–337.