Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied & Preventive Psychology 11 (2004) 47–51
APPLIED & PREVENTIVE PSYCHOLOGY
http://www.elsevier.com/locate/app

Commentary

Taking theoretical risks in a world of directional predictions

Scott O. Lilienfeld

Department of Psychology, Emory University, Room 206, Atlanta, GA 30322, USA

Abstract

As Meehl (1978) observed, one major reason for the painfully slow progress of soft psychology is our field's heavy reliance on directional predictions, which leads to feeble tests of theories. In the present commentary, I outline six ways of taking theoretical risks in a world of directional predictions: (1) constructive replication of findings, (2) generating multiple maximally independent predictions, (3) adherence to the total evidence rule, (4) discriminant validation of measures, (5) incremental validation of measures, and (6) according attention to higher-order personality dimensions.

Keywords: Significance testing; Replication; Discriminant validity; Incremental validity; Higher-order dimensions

Reading anything authored by Paul Meehl is a humbling experience. Re-reading Meehl's classic "Two Knights" paper, which has long been of one my favorites, reminded me of Meehl's remarkable breadth of knowledge and incisiveness of thinking. Finding something substantive to add to this article, which I regard as one of the most significant (clinically, not statistically, that is) articles in psychology in the past several decades (see Rorer & Widiger, 1983, for a similar view), is a daunting task.

In the brief space I have available, I've elected to elaborate on one central issue Meehl (1978) raised. As Meehl observed, one of the principal reasons for the excruciatingly slow progress of soft psychology is our field's collective failure to subject theories to stringent theoretical risks. Most predictions generated by soft psychological theories are merely directional: they forecast nothing more than the direction of a statistical effect (e.g., men will be more aggressive than women, extraverts will exhibit lower resting heart rates than introverts). The problem is that in the domain of individual differences, the null hypothesis is virtually always false (see Section 4). Hence, as statistical power approaches 1.0, the probability that our hypotheses will be corroborated approaches 50% even if the theory generating these hypotheses is entirely without merit. This is a profoundly troubling state of affairs.

Meehl (1978) contended that one means of circumventing our longstanding tendency to rely on feeble tests of theories is move beyond directional hypotheses by positing more specific (read: risky) quantitative relationships. Such hypotheses include predictions concerning the precise value of a parameter ("point predictions"), the rank-order of numerical differences, the ranges of expected values (see also Lykken, 1968, for a discussion of "range predictions"), and the function forms of statistical associations. Yet as Meehl acknowledged, doing so is far easier said than done. Toward the conclusion of his article, Meehl mused pessimistically that "It may be that the nature of the subject matter in most of personality and social psychology is inherently incapable of permitting theories with sufficient conceptual power (especially mathematical development) to yield the kinds of strong refuters expected by Popperians, Bayesians, and unphilosophical scientists in developed fields like chemistry" (p. 829). That is, in most domains of soft psychology, we may be trapped—at least for the foreseeable future—in a world of directional predictions.

Given that most soft psychological theories are not nearly as powerful as those in the hard sciences (but see Wilson, 1998, for the contention that the social sciences, not physics or chemistry, are the genuinely "hard" sciences), is there anything we can do to "make our theories work harder," as my friend Allan Harkness (in press) likes to say?

Yes we can, and even in a world of directional predictions there are ways of subjecting our theories to stronger theoretical risks. Here, I'll outline six such approaches, with the full realization that (a) these six approaches are merely a selective sampling of suggestions for how to

E-mail address: slilien@emory.edu (S.O. Lilienfeld).

0962-1849/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.appsy.2004.02.008

generate riskier tests of theories and (b) nothing I will say here has not been said before, and been said more eloquently, by others. Nevertheless, by compiling a user-friendly list of methodological approaches for living dangerously in hypothesis-testing world fraught with safety, I hope to encourage the next generation of psychological researchers to take more frequent risks and therefore become better scientists. For the sake of brevity, I will omit approaches relevant to treatment outcome research, such as dismantling designs (Kazdin, 2003), as well as familiar approaches that permit the identification of suppressor (Conger & Jackson, 1972) and moderator (Jaccard, Turrisi, & Wan, 1990) variables. Nevertheless, all of these approaches and several others that I will not mention permit at least somewhat risky tests of theoretically informed hypotheses. In several of the examples I present, I draw from research on psychopathic personality simply because it is a content area with which I am familiar. But readers should bear in mind that my points apply to virtually all research domains within soft psychology.

1. Constructive replication

As Lykken (1968) noted, one means of minimizing the difficulties associated with an exclusive reliance on significance testing of directional predictions is to insist on replication using alternative measures of the same constructs examined by the original investigators, a procedure he termed "constructive replication." Constructive replication typically affords riskier tests of the investigator's predictions than does "operational replication," which relies on the same measures used by the original investigators, because it minimizes the problem of systematic error arising from the use of measures sharing method variance (see Messick, 1995, for a discussion of "construct irrelevant variance").

Of course, the primary reinforcement contingencies in our field militate against constructive replications. Most journal editors place a premium on novelty. I once co-authored an article that was initially rejected by a prominent journal (fortunately, it was eventually published in that journal following revisions) on the grounds that it was "only" a replication of a study published previously in that journal. Putting aside the fact that we replicated the earlier findings using different measures of cognate constructs, the fact that we managed to constructively replicate other investigators' findings should be accorded considerable weight. Given the dismayingly low replication rates in most social scientific disciplines (Cronbach, 1975), one could argue that replications should generally be accorded more, not less, weight in psychological journals than are original findings. Fortunately, a few high-quality journals, such as the *Journal of Personality and Social Psychology* and the *Journal of Experimental Psychology: General*, encourage the publication of articles featuring multiple studies that constructively replicate and extend authors' initial findings.

2. Multiple maximally independent predictions

One of the best ways of subjecting theories to risky tests is to generate multiple independent directional predictions (again see Lykken, 1968, for a discussion of "multiple corroboration"). The more such divergent predictions are corroborated, the greater should be the investigator's confidence in the theory that spawned them. By capitalizing on the "heterogeneity of irrelevancies" (see Shadish, Cook, & Campbell, 2002), that is, the presence of independent and non-overlapping methodological limitations across studies, researchers can enhance the likelihood that the corroboration of multiple directional predictions is not merely a function of shared errors (e.g., similar confounds, sampling biases) across investigations.

As an example, Lykken (1968) described the results of his classic dissertation (Lykken, 1957), which yielded three independent corroborations of the theory that psychopathy stems from a deficit in fearfulness ("a low fear IQ") using three different procedures—a self-report measure of fearfulness, a classical conditioning task involving aversive stimuli, and a passive avoidance task. Although Lykken's fearlessness theory is not easily reconciled with all of the data on psychopathy (Newman & Brinkley, 1997), the fact that it has been corroborated using markedly different methodological approaches (see Patrick, Bradley, & Lang, 1993, for another corroboration of this theory using the fear-potentiated startle paradigm) lends it a well earned cachet of scientific respectability.

3. Adherence to the total evidence rule

I am surprised by how frequently psychological investigators pursue the following methodological strategy: they advance Theory X to explain Phenomenon Y (e.g., a specific disorder), design studies to test Theory X, yet all the while pay virtually no heed to a host of previous findings bearing on Phenomenon Y. Such a methodological strategy is inherently flawed, as it neglects the obvious problem that some well replicated findings bearing on Phenomenon Y may not be explicable by Theory X. Such researchers are running afoul of Carnap's (1962) "total evidence rule," the dictum that researchers must account for all of the scientifically credible evidence, not merely the evidence cherry-picked to suit their theoretical views.

For example, one of the earliest and best replicated findings in the psychopathy literature is that psychopaths exhibit weak skin conductance responses in anticipation of aversive stimuli in classical and quasi-conditioning paradigms (e.g., Hare, 1965; Lykken, 1957). Presumably, they are less afraid of impending danger than are non-psychopaths. Although a number of sophisticated models of psychopathy, such as Newman's response modulation model (see Patterson & Newman, 1993) and Kosson's (1998) left hemisphere deficit hypothesis model, have been developed

since these consistent findings appeared in print, it remains to be seen whether these models can accommodate the earlier findings. For example, Newman's model posits that psychopaths' problematic traits and behaviors stem from their failure to attend to extraneous stimuli, including punishment, once they are engaged in a dominant response set. But is not clear whether this model can account for psychopaths' deficits in aversive classical conditioning paradigms (e.g., Lykken, 1957). Only if such models can be shown to account for all credible evidence will they themselves be credible (but see Newman & Brinkley, 1997, for a critique of Lykken's fearlessness model on similar grounds).

4. Discriminant validation

In the world of individual differences, virtually everything is at least somewhat correlated with everything else. That's largely because the number of input variables into virtually all individual difference variables is enormous and of varying weight, so that for two individual difference variables to be entirely uncorrelated in the population it would be necessary for all of their input variables, not to mention their associated beta weights, to cancel out exactly. The "crud factor," or low to modest level of ambient covariance among virtually all individual difference dimensions (Lykken, 1991), renders an overreliance on convergent validation problematic, not to mention decidedly non-risky, as a methodological strategy. Given sufficient statistical power, it is virtually inevitable that our newly developed measure X will correlate at least slightly with a convergent validity target variable. Moreover, the odds approach 50% that measure X will exhibit "convergent validity" with this target variable if our measure is reliable but entirely invalid as an indicator of its intended construct.

This problem implies that the discriminant validation of measures should typically be at least as important, if not more, than convergent validation. That's because a measure can be invalidated by correlations with other measures that are too high (Campbell, 1960). Yet, many articles reporting on the development of novel measures place negligible emphasis on discriminant validity. Campbell and Fiske (1959) noted this void over 4 decades ago, but it does not appear to have filled noticeably since then.

This problem is exacerbated by the frequent tendency of authors to use the term "discriminant validation" to refer to the demonstration of predicted group differences (such as between a psychopathological group and a normal comparison group; e.g., see Canals, Carbajo, & Fernandez-Ballart, 2002; Gripshover & Dacey, 1994) on a measure. In fact, such a finding demonstrates only convergent validity, because it reveals a predicted point-biserial correlation between a measure and the presence or absence of a grouping variable. Discriminant validity refers to the predicted absence of a strong statistical association, not to its presence.

5. Incremental validation

The decision to construct a new psychological measure virtually always hinges on a miniature implicit hypothesis, namely that this measure is not redundant with extant measures (the lone exception being the case in which one develops a more economical measure in the hopes that it is redundant with an extant measure). Otherwise, why bother constructing it? Yet, surprisingly few developers of new measures take the trouble to demonstrate these measures' incremental validity (Meehl, 1959; Sechrest, 1963), that is, the extent to which they contribute surplus information not available from other instruments (Garb, 1984; Hunsley & Meyer, 2003). Haynes and Lench (2003) located 298 manuscripts submitted over a 4.5-year period to *Psychological Assessment* dealing with the development and validation of a new measure. Only 26 reported data on incremental validity.

For example, there is little evidence that the most extensively validated measure of psychopathy, the Psychopathy Checklist-Revised (PCL-R; Hare, 1991), possesses incremental validity in the detection or prediction of important variables (e.g., recidivism, performance on laboratory measures) beyond more easily administered measures, such as self-report measures (Lilienfeld, 1998). This fact is surprising given that the PCL-R requires an extensive interview (itself requiring formal training) and review of file information. In one of the few investigations of this question, Edens, Poythress, and Lilienfeld (1999) found that neither the PCL-R nor the Psychopathic Personality Inventory (Lilienfeld, 1996), a questionnaire measure of psychopathy, exhibited incremental validity beyond the other measure in predicting prisoner disciplinary infractions, although the relatively low zero-order correlations of both measures with these infractions rendered this study a less than optimal test of incremental validity.

6. Attention to higher-order dimensions

One of the most serious—and common—methodological and conceptual errors in psychopathology and personality research is the neglect of the potential influence of higher-order dimensions that cut across many or most measures. As Watson, Clark, and Harkness (1994) observed, higher-order dimensions often offer competing explanations for hypotheses involving lower-order dimensions. Nevertheless, many investigators neglect to incorporate measures of higher-order dimensions in their studies, perhaps because such measures can complicate the interpretation of their findings and place constraints on their conclusions.

Take the higher-order dimension of negative emotionality (NE; Tellegen, 1982; Watson & Clark, 1984), a pervasive disposition to experience aversive emotions of many kinds, including anxiety, guilt, anger, impatience, and mistrust. This generalized maladjustment dimension courses through so many self-report measures of psychopathology

that it is difficult to develop a questionnaire that is not saturated with it (Finney, 1985; Tellegen, 1985). Yet many investigators routinely administer measures of lower-order constructs without concurrently administering measures of NE, but then draw theoretical inferences specific to these lower-order constructs.

For example, in the life events literature, many investigators do not sufficiently appreciate the extent to which responses to relatively mild life events (e.g., "daily hassles;" Lazarus, 1984) are largely attributable to the contaminating influence of NE (Depue & Monroe, 1986). The heavy saturation of life events items with NE may in part reflect the tendency of individuals with elevated levels of NE to experience minor irritants as extremely stressful, and as well as their tendency to elicit antagonistic responses from others (Watson & Clark, 1984). Hence, many of the widely reported associations between life events scales and measures of psychological distress may merely reflect the saturation of both set of instruments with NE. To take another example, one is left to wonder how many of the inferences regarding the relation between the Type A construct and self-reported adverse health outcomes are attributable to NE rather than to the specific features of the Type A construct, such as hostility and time urgency, most of which are shared with NE (e.g., Bruck & Allen, 2003). NE may contribute to poor self-reported health outcomes by influencing the interpretation of ambiguous physical symptoms (Watson & Pennebaker, 1989).

Physicist and Nobel Laureate Feynman (1985) reminded us that the essence of science is bending over backwards to prove ourselves wrong. Paul Meehl (who struck up a correspondence with Feynman during the last years of the great physicist's life) in turn reminded us that the best way to accomplish this goal is to lay our most cherished theoretical conjectures on the line. In so doing we risk proving ourselves wrong, but we at least stand a chance of learning something in the process.

References

Bruck, C. S., & Allen, T. D. (2003). The relationship between big five personality traits, negative affectivity, Type A behavior, and work-family conflict. *Journal of Vocational Behavior, 63*, 457–472.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, and discriminant validity. *American Psychologist, 15*, 546–553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Canals, J., Carbajo, G., & Fernandez-Ballart, J. (2002). Discriminant validity of the Eating Attitudes Test according to American Psychiatric Association and World Health Organization criteria of eating disorders. *Psychological Reports, 91*, 1052–1056.

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press.

Conger, A. J., & Jackson, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement, 32*, 579–599.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 117–127.

Depue, R. A., & Monroe, S. M. (1986). Conceptualization and measurement of human disorder in life stress research: The problem of chronic disturbance. *Psychological Bulletin, 99*, 36–51.

Edens, J. F., Poythress, N. G., & Lilienfeld, S. O. (1999). Identifying inmates at risk for disciplinary infractions: A comparison of two measures of psychopathy. *Behavioral Sciences and the Law, 17*, 435–443.

Feynman, R. P. (1985). *Surely you're joking, Mr. Feynman*! *Adventures of a curious character*. New York: W.W. Norton and Company.

Finney, J. C. (1985). Anxiety: Its measurement by objective personality tests and self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 645–673). Hillsdale, NJ: Erlbaum.

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review, 4*, 641–655.

Gripshover, D. L., & Dacey, C. M. (1994). Discriminative validity of the MacAndrew Scale in settings with a high base rate of substance abuse. *Journal of Studies of Alcohol, 55*, 303–308.

Hare, R. D. (1965). Temporal gradient of fear arousal in psychopaths. *Journal of Abnormal Psychology, 70*, 367–370.

Hare, R. D. (1991). *Manual for the Psychopathy Checklist-Revised*. Toronto: Multi-Health Systems.

Harkness, A. R. (in press). Essential Paul Meehl lessons for personality assessment psychology. *Journal of Clinical Psychology*.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new assessment measures. *Psychological Assessment, 15*, 456–466.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.

Kazdin, A. E. (2003). *Methodological issues and strategies in clinical research*. Washington, DC: American Psychological Association.

Kosson, D. S. (1998). Divided visual attention in psychopathic and nonpsychopathic offenders. *Personality and Individual Differences, 24*, 373–391.

Lazarus, R. S. (1984). Puzzles in the study of daily hassles. *Journal of Behavioral Medicine, 7*, 375–389.

Lilienfeld, S. O. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment, 66*, 488–524.

Lilienfeld, S. O. (1998). Methodological advances and developments in the assessment of psychopathy. *Behaviour Research and Therapy, 36*, 99–125.

Lykken, D. T. (1957). A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology, 55*, 6–10.

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

Lykken, D. T. (1991). What's wrong with psychology, anyway? In D. Cichetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (Vol. 1., pp. 3–39). Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology, 13*, 102–128.

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.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Newman, J. P., & Brinkley, C. A. (1997). Reconsidering the low-fear explanation for primary psychopathy. *Psychological Inquiry, 8*, 236–244.

Patrick, C. J., Bradley, M. M., & Lang, P. J. (1993). Emotion in the criminal psychopath: Startle reflex modulation. *Journal of Abnormal Psychology, 102*, 89–92.

Patterson, C. M., & Newman, J. P. (1993). Reflectivity and learning from aversive events: Toward a psychological mechanism for the syndromes of disinhibition. *Psychological Review, 100*, 716–736.

Rorer, L. G., & Widiger, T. A. (1983). Personality structure and assessment. *Annual Review of Psychology, 34*, 431–463.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23*, 153–158.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Tellegen, A. (1982). Brief Manual for the Multidimensional Personality Questionnaire. Unpublished manuscript, University of Minnesota.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma, & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681–706). Hillsdale, NJ: Erlbaum.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin, 96*, 465–490.

Watson, D., Clark, L. A., & Harkness, A. R. (1994). Structures of personality and their relevance to psychopathology. *Journal of Abnormal Psychology, 103*, 18–31.

Watson, D., & Pennebaker, J. W. (1989). Health complaints, stress, and distress: Exploring the central role of negative affectivity. *Psychological Review, 96*, 234–254.

Wilson, E. O. (1998). *Consilience: The unity of knowledge*. New York: Vintage Books.