Meehl, P. E. (2006). The power of quantitative thinking. In N.G. Waller, L.J. Yonce, W.M. Grove, D. Faust, & M.F. Lenzenweger (Eds.), *A Paul Meehl Reader: Essays on the practice of scientific psychology* (pp. 433–444). Mahwah, NJ: Erlbaum.

18 The Power of Quantitative Thinking

At age 14, having been fascinated by Karl Menninger's book, *The Human Mind*, I decided to become a psychoanalyst rather than a lawyer. In the late 1930s, the Minnesota psychology department was not favorably inclined toward Freud's teachings. Theoreticians William T. Heron and B. F. Skinner were archbehaviorists, and the applied psychologists John G. Darley, Starke R. Hathaway, and Donald G. Paterson were hyper-operational psychometricians. John Harding left Minnesota to get his PhD at Harvard, and he described our department as "the hotbed of dust-bowl empiricism." Professor Charles Bird, who taught an excellent course in abnormal psychology, grudgingly allowed for the existence of unconscious mental processes, and, like my mentor Hathaway, did think well of the mechanisms of defense. It is exactly 100 years since Freud abandoned the pressure technique in favor of free association; he considered this, especially as applied to the interpretation of dreams, to be his big methodological discovery. But we still do not know with any scientific confidence whether Freud's doctrines are 10%, 50%, or 90% correct. My colleague, David T. Lykken, whose analyst was the same as mine, has become quite skeptical about the whole business, although I note that, like me, he has a picture of Freud on his office wall. A former Skinnerian colleague rejected all of it, and in class lectures habitually referred to its creator as "Dr. Fraud." How are we to explain this amazing episode in the history of science, a system of ideas which has had such a profound and pervasive effect on the thinking of almost all educated persons but, after a century of research and practice, is still in such doubtful status?

The standard explanations of my experimental colleagues, despite elements of truth, are inadequate, and I think wrong-headed. It is not the alleged difficulties of observation because, as Skinner pointed out in his classic *Behavior of Organisms*, compared with most sciences the phenomena of behavior are relatively macroscopic and slow, and they can be objectively recorded for subsequent analysis at leisure. The problem is *not* recording the events but slicing the pie, of classifying and numerifying in fruitful ways. Nor does the problem lie in the novel use of mentalistic language, as when we speak of an unconscious fantasy. Nor is there anything intrinsically objectionable about the theoretical constructs introduced. Nor is it because we cannot perform experiments, a criterion which would eliminate astronomy, geology, comparative anatomy, and other legitimate and fairly respectable sciences.

After many years of reflection and psychoanalytic practice, I am firmly persuaded that the difficulty is *epistemological* and, more specifically, *statistical*. It lies in the lack of an adequate quantification. Lord Kelvin said that a subject matter is scientific to the extent that it is quantitative. Quantification is one of

---

This chapter was delivered as a speech when Meehl received the James McKeen Cattell Fellow Award from the American Psychological Society, Washington, D.C., May 23, 1998.

the main features of the scientific method (which does exist, despite the obscurantists' claim). In the psychoanalytic session, the basic epistemic problem is that the human mind is ingenious and able to explain almost anything. We confront a situation in which the application of a purported general theory to a particularistic hypothesis is markedly underdetermined. Speaking roughly, we have far too many unknowns in relation to the number of equations. A simple model of a psychoanalytic episode, such as a parapraxis or the associations to a dream, would be to classify the analyst's interpretation in terms of needs, objects, and defenses. Some 20 Murray needs and 20 defenses yields 400 need-defense combinations; adding 40 objects gives us 16,000 configurations. With a little ingenuity and plenty of looseness, explanations abound. In a personality theory seminar, after I had explained a particular dream interpretation, Gardner Lindzey challenged my anti-Freudian mentor, Hathaway, to say what was objectionable about it. Hathaway's response was, "Well, it's a free country. If Meehl thinks it helps him to see the thing that way, I have no objection." Not a satisfactory scientific situation!

There are some sessions in which what Freud calls the "red thread" running through the associations is so easily discerned that one has high, although still subjective, confidence in its identity. There are others, more typical, in which repeated allusions—sometimes obvious, often subtle and debatable—occur, but much of the material is sawdust, interpretively speaking. And then, alas, there are sessions in which neither analyst nor analysand can discern much of anything that hangs together.

Any documentary, nonexperimental discipline in which this mixture of clear and unclear episodes typically occurs presents a terrible statistical problem to a skeptical mind. Archeology, history, biography, literary criticism, and paleontology share this epistemic curse with psychoanalysis, so that when a disillusioned training analyst like Alan Stone says that the future of psychoanalysis lies in our viewing it as we do fields like literary criticism, this is hardly reassuring. Freud's famous analogy to a jigsaw puzzle is a false analogy because you know for sure when you've solved the jigsaw puzzle: there are no holes, no unused pieces, and an unmistakably meaningful picture. Even if we had solved the problem of statistically proving a nonchance *thematic pattern*, we would still be confronted with philosopher Grünbaum's criticism of one of my psychoanalytic papers, that this does not tell us about the *inferred latent causality* guiding the associations.

There are *three kinds* of quantification in science. The first is at the level of measurement, the numerification of observations. The second is summary statistics of the numerified observations, such as means, variances, correlations, and curve fitting. The third is mathematicization of the *theoretical entities*. A good science has all three, connected in nonarbitrary ways. In the social sciences, one often sees both quantification of the observations and mathematicization of the theoretical concepts, but only tenuous linkages between the two. Economics, despite its pretensions of being the queen of the social sciences, is chronically in

that situation. Some psychologists who write articles for *Psychometrika* teeming with inverse matrices and Jacobians are rather like the economists.

In my 1954/1996 book on prediction, I discussed two different uses of statistics, which involve the interpretive text rather than the formalism. In the *discriminative-validating* use, one goes very little beyond the data, and the only theory involved is the theory of probability that we get from the statisticians. This use is unavoidable for anyone who makes quantitative claims, and it is remarkable how many anti-quantitative psychologists do not understand that simple point. Anytime one speaks of something being typical, or rare, or of an influence being great or small, or of two things being related—those are *intrinsically statistical claims*, although they are in ordinary English or social science jargon; and there is no known method of testing a statistical claim except to compute statistics. The other use, which I call the *structural-analytic* use (exemplified by such procedures as factor analysis, multidimensional scaling, or my taxometric method), goes beyond the statisticizing of the observations and involves inference problems about theoretical entities for which the mathematician provides no algorithm. Quantitative thinking is not the same as mathematical knowledge, although they tend to be correlated. Law students have high quantitative talent, despite usually having little mathematical knowledge. Let me briefly examine some examples of progress and problems in quantification.

Verbal definitions of the intelligence concept have never been adequate or commanded consensus. Carroll's (1993) *Human Cognitive Abilities* and Jensen's (1998) *The g Factor* (books which will be the definitive treatises on the subject for many years to come) essentially solve the problem. Development of more sophisticated factor analytic methods than Spearman or Thurstone had makes it clear that there is a *g* factor, that it is manifested in either omnibus IQ tests or elementary cognitive tasks, that it is strongly hereditary, and that its influence permeates all areas of competence in human life. What remains is to find out what microanatomic or biochemical features of the brain are involved in the hereditable component of *g*. A century of research—more than that if we start with Galton—has resulted in a triumph of scientific psychology, the footdraggers being either uninformed, deficient in quantitative reasoning, or impaired by political correctness.

The clinical-statistical problem ("What is the optimal method of combining data for predictive purposes?") has also been solved, although most clinicians have not caught on yet. A meta-analysis of 136 research comparisons of informal with algorithmic data combination, conducted by Will Grove, has settled this question. I do not know of any controversy in the social sciences in which the evidence is so massive, diverse, and consistent. It is a sad commentary on the scholarly habits of our profession that many textbooks (and even encyclopedias) persist in saying that the question is still open. I suppose they would be saying that if we had 1360 studies.

A typical reaction is to say benignly, "Oh, well, this is a spurious issue. I am a *clinician*, I use both methods." This sounds amicable, tolerant, and even-handed, but it is actually stupid. If the regression equation or actuarial table predicts

Jones will wash out of flight training, and my impressionistic judgment says he will succeed, how can I "use both"? Do I cut Jones in half, as suggested by King Solomon? Admission to flight training is a dichotomous administrative act, we admit or reject. When statistical and nonstatistical predictions collide, as they often do, we rarely have a "compromise option" lying between them. Suppose the equation says Patient X will benefit from shock therapy, but the psychiatrist thinks not. Do we compromise by using half as many volts A.C. as the potential needed to induce a cerebral storm? If the last step in decision making can countervail the equation by human judgment, the procedure is "clinical," for purposes of this discussion. The melioristic response is sheer denial of a real, concrete, unavoidable decision problem, and I note that it does not involve ignorance of higher mathematics, but a simple inability to think clearly.

In my book (Meehl 1954/1996), I considered situations in which a special, rare fact is present that completely countervails a strong statistical prediction. For example, knowing Professor X's values and interests and his track record of movie attendance, we find from our actuarial table or discriminant function that there is a probability $p = .84$ of his going out to a certain movie Saturday night. Then we learn that he has a broken thigh and is in a hip cast, which trumps all our actuarial data, and drops the probability to zero. Some critics of my book forget that Meehl's broken leg case was examined there, and that the obvious existence of such special cases does not tell us whether, and how often, a clinician should cancel the deliverances of the equation because she thinks it is a broken leg case. The most common objection to proceeding actuarially is to say, "Well, but special cases arise where I see something that's important that the equation does not take account of."

Let us do a little ninth-grade algebra: In each individual case, the clinician's prediction—response to shock therapy, success in dental school, survival in aircrew training, parole violation—either agrees with the equation or not. For the subset of *agreements*, the hit rate for the equation and the clinician will be identical. Considering the subset in which the clinician's prediction *disagrees* with the equation, is he more likely to be right than wrong? If he is more likely to be right, his hit rate will exceed that of the equation for the disagreement subset. It follows as the night the day that his *overall* hit rate will exceed that of the equation. But the *empirical research shows that it does not*. Conclusion, for anyone who can add: The clinician is not as good at spotting broken leg cases as he thinks he is. Failure to understand this simple arithmetical argument reflects a grave defect in quantitative reasoning ability.

But there is an interesting follow-up consideration that is rarely discussed in the literature by either side of the dispute. Suppose an educable, rational clinician, properly humbled by reading the comparison research, decides to raise the standards for calling a broken leg. A rational person knows that most social and psychological attributes are not as reliably recognized as broken legs; and that even when present as perceived, they do not have the countervailing causal efficacy that broken legs do in reducing mobility. So, our psychologist regularly reduces the broken leg countervailing calls. Will this improve things? Not neces-

sarily. The research suggests that the clinician's main problems in competing with the equation are twofold: first, assigning nonoptimal weights to factors that are in the equation with weights nearer the optimal; and second, applying those nonoptimal weights inconsistently. If these error sources preponderate over mistaken countervailings, the clinician may do better by calling a few broken legs, so that the statistician's friendly advice might make her worse instead of better. That, of course, is no answer to the proactuarial argument unless we know more details about the empirical parameters.

Now, I come to what could be described as a scientific scandal, the significance test controversy. It has two major components, either of which should make us nervous. First is the problem of the power function. In 1962, Jacob Cohen exhaustively sampled one year's issues of the *Journal of Abnormal Psychology* and found that the majority of investigations had grossly inadequate statistical power to detect real differences. I am not aware that anybody challenged his facts or denied their implications. What was the effect of this classic paper? Twenty-seven years later, Sedlmeier and Gigerenzer exhaustively sampled a year's issues of the journal, and they found that the statistical power of the studies had undergone a slight *decline*! This tells us there is something terribly wrong with the intellectual discipline of psychologists. Given the rejection rate of 85%, I assume that the authors of these defective papers are in the upper decile of research competence. What cognitive shape must most psychologists be in? It is inconceivable that such a thing could happen in a strong science like chemistry. If a chemist published a paper showing that the litmus test worked less than half the time, can we imagine that almost three decades later chemists would still be publishing articles using it, without scruple or comment? This reveals a deficiency in quantitative thinking of grave proportions.

The other side of the coin is as bad or worse. After adumbrations in the 1940s and 1950s by statisticians Berkson (1942) and Hogben (1957/1968), in the 1960s several psychologists pointed out the weakness of null-hypothesis *refutation* as a means of corroborating psychological theories. I was one of the first (Meehl, 1967a), preceded by Rozeboom (1960) and Bakan (1966), followed by Lykken (1968), Carver (1978), and Jacob Cohen (1994) again (addressing himself to the other side of the coin). It took 30 years of such unanswered criticism before the American Psychological Association (APA) woke up to the fact that there might be a problem here, and appointed a committee to examine it. This is a gorgeous example of the resistance to scientific discovery as described by sociologists Barber (1961), Merton (1973), and others.

I do not wish to be invidious, but I am afraid that the APA committee has labored to bring forth a mouse. The report (Wilkinson & Task Force on Statistical Inference, 1999) reads like a politician's "blue-ribbon" committee, coming out in favor of motherhood, the flag, and apple pie; and it has no teeth in it. It does not require or forbid anything, including the most irrational current practices. I was one of the four outside experts, named as consultants—the others being Fred Mosteller, John Tukey, and Lee Cronbach—and I would be curious

to know whether the committee paid as little attention to the other three as they did to me. Let me explain the solution to this long-standing problem.

The first thing to be clear about is the distinction between a substantive scientific theory, *T*, and the statistical hypothesis, say, *H**, that flows from it. Hardly any statistics book has so much as a single sentence making this distinction, when pedagogically it deserves a page or two. If I were to write a statistics book, it would get a whole chapter. This egregious pedagogical error seduces psychologists into thinking that if they have strongly refuted a null hypothesis H_0 at level α, thereby strongly corroborating its statistical alternative *H**, somehow the big value of $(1 - \alpha)$ strongly corroborates theory *T*, which it almost never does. I explained this clearly in 1967, but without effect on the profession, including the APA committee. In developed sciences such as physics, astronomy, chemistry, and genetics, the semantic difference between *T* and *H** is present, but it does less damage than it does to us. In those disciplines, the *strong use* of significance tests prevails: the theory is strong enough to predict a numerical point value or a narrow interval, and the scientist subjects the theory to the danger of falsification if the observed values differ significantly from the predicted. Most psychologists do not realize that's what Karl Pearson (1895) had in mind when he invented χ^2, as the title of his classic article shows. Since in social science everything is correlated with everything (Lykken's crud factor), whether one gets apparent support for a false theory by refuting H_0 depends solely upon the power function. Associate a theory—*true or false*—with any arbitrary pair of observational variables. With perfect power, you have a 50:50 chance of refuting H_0 in a "predicted" direction. In a domain with positive manifold (ability, psychopathology, social attitudes, achievement), the theory is nearly certain to be confirmed, despite its having no logical relation to the facts. The problem here is only partly mathematical (that's Cohen's side of the coin); the other side is a matter of freshman logic and epistemology, not statistics. You might hope that Cohen's complaint and mine, tending oppositely, could somehow countervail each other, so all is well. Not so. High statistical power plus crud factor pseudo-confirms false theories; low power pseudo-refutes true theories. The net effect is that the empirical success rates of true and false theories are brought closer together, and I have shown on reasonable assumptions that their box scores may differ by as little as 10 or 15% [see chap. 19].

Anyone who seriously wants to do something about the scandal of null-hypothesis refuting would lay down certain rules that editors were bound to follow. I have such a set of rules that I will almost guarantee will solve the problem without improperly censoring scholars.

First, almost all situations in which a significance test can be properly employed permit setting up of a confidence interval. That should come first.

Second, if significance tests of the weak kind are done, the author may state in a sentence explicitly what it *means*, but the use of the cancerous *term* 'significant' is prohibited. If this troubles you as censoring freedom of scholarly expression, I remind you that a 37-page chapter of the APA publication manual consists of such injunctions and forbiddings about the use of language, many of

them involving matters less scientifically important than this one. All statistics texts and, I daresay, all lecturers for two generations, have explained that the word 'significant' here has a narrow, specific, mathematical meaning and does not impute practical or theoretical importance. Since these routinely reiterated statements have not cured the disease, it is appropriate to adopt a more drastic therapy. All you forbid is the use of this malignant, misleading expression. You do not forbid a sentence that says exactly the same thing.

Third, if comparisons are made between results that reach α and those that do not, the author must present the statistical power. Failing to do so is not a minor weakness, it makes the manuscript unacceptable. Leaving out the value of the power function is as bad scientific reporting as not telling the sample size or where you got your subjects or which measuring instrument you used.

Fourth, an appropriate measure of overlap must be presented. The reader should not be required to try to figure that out in his head. Donald Paterson taught us to present what proportion of the one group reached or exceeded the 10th, 25th, 50th, 75th, and 90th percentile of the other group. Needless to say, Cohen's effect size, or κ, or other measures of association should be required when appropriate.

Finally, there must be an explicit statement to the effect that since the statistical hypothesis H^* (that has been indirectly confirmed by refuting H_0) is not semantically equivalent to the substantive theory T being studied, the confidence in H^* cannot be transferred to a similar confidence in T. Some complain that here I would be imposing Meehl's or Popper's philosophy of science. Not so. I am only imposing something that one learns in freshman logic about the third figure of the hypothetical syllogism.

In my 1967 paper, I was overly Popperian, and have been properly corrected by Serlin and Lapsley (1985) so as to emphasize what they call the "Good Enough Principle." The quantitative implementing of the Good Enough Principle is the most important methodological task for those who recognize the inadequacy of the present mindless proving that the A's differ from the B's. It will be important for us to listen not only to creative statisticians, but to logicians and philosophers of science. Because of my emphasis upon severe tests, which refuting H_0 is not and cannot become, some perceive me as a staunch disciple of Sir Karl Popper. This is incorrect. I count approximately 16 major theses which Popper defended throughout his life, and I agree with only three of them.[31]

Part of the problem of shifting to severe tests is our pessimism about the *possibility* of concocting strong theories capable of generating point or narrow interval predictions. Combining this pessimism with the almost indestructible optimism about the present mode of refuting H_0 rather than refuting the consequences of a theory, we have an impasse, which I do not know how to surmount. Perhaps exposure to the history of the other sciences would help. Psychologists seem to think that you cannot make stronger theories unless they are as strong as physicists have, and that physicists have always been able to derive precise point predictions. This is not the case, especially in the early stages of theory development. It is sometimes possible to derive from fairly weak theories statements

(about rank order, or about function forms without providing parameters, and the like) that can subject the theory to severe tests. Consider Wien's Law concerning the radiation in a black body, out of which, after Planck's epoch-making substitution of a sum for an integral (1900), grew the powerful branch of physics we know as quantum mechanics. Based on an accepted theory of the furnace wall particles as tiny harmonic oscillators, Wien derived a formula predicting that the proportion of radiation emitted from a peephole in the black body at each specified wavelength λ should equal *some function* of the product of λ and the Kelvin temperature, divided by the 5th power of the wavelength, and that this function should not depend upon the temperature. The theory was too weak to derive the function with its parameters, or even the function *form* (a parabola, a hyperbola, or whatever). If one plots the product $\lambda^5 f(\lambda T)$ at temperatures of 1000, 2000, 3000 degrees Kelvin, a French curve can draw 3 beautiful smooth graphs going precisely through the points and coinciding at all three temperatures—but you still do not know what the function is. I believe there would be many examples of this, even in such soft areas as my own field of psychopathology, if we had the wit and the determination to try it. But it will take people with both the talent for quantitative thinking and the mathematical knowledge to be a little creative. And by 'mathematics' I do not mean learning how to look up the probability of a χ^2 in the back of the statistics book. I see no excuse, given the history of the other sciences and the relative rate of progress of various fields of psychology, for the abysmally poor mathematical education that we require of our students. I shall say more about this later.

The heated controversy over path analysis exemplifies the difference between mathematical knowledge and sound quantitative reasoning. Some experts think it is the royal road to untangling causal relations in complex correlational systems; others consider it nearly useless. I am no expert, but I do know there are widespread abuses. I have seen analyses in which a dozen plausible path diagrams are riffled through, the *least poor* fit selected, and strong conclusions drawn about weighted causality of drug abuse or delinquency. This is algorithmic *ad-hoc*kery. Maximizing a likelihood function by zeroing a partial derivative is here a mathematical fig leaf covering a gross methodological indecency. We cannot blame LISREL's inventors Jöreskog & Sörbom (2001) for these abuses, against which they issued loud and clear warnings. We must amend or supplement LISREL so as to provide strong Popperian *risk*. [A solution to the problem was subsequently proposed by Meehl and Waller (2002) and Waller and Meehl (2002).]

Finally, consider the controversy about *DSM*, the psychiatrists' classification of mental disorders. To everybody, both our trade union and the MDs, it is somewhat unsatisfactory. In the eyes of some competent scholarly practitioners, parts of it are laughable. Truth by committee is never scientifically satisfactory, although for political and educational reasons, it is sometimes unavoidable. Whatever the defects of detail, the first thing wrong with *DSM* is its espousal of a simplistic operationism in defining constructs. No philosopher of science or logician has believed in simple operationism since 1936, at the latest. Contrary

to what some keep teaching in introductory psychology, not all scientific concepts are operationally defined, and efforts to do this for all of them turn out to be fraudulent. The meaning of most scientific terms (especially in powerful sciences like astronomy, physics, chemistry, and genetics) is given by their role in the postulated nomological network, as pointed out by Cronbach and Meehl in 1955. The nodes are the theoretical concepts, and the strands are the relations postulated between them. When the network is incomplete, and when the strands of the net are, as in the soft sciences, stochastic rather than nomological, then there is a sense in which the concepts are only partially defined. The incomplete and stochastic character of the postulated law network gives rise to philosopher Arthur Pap's (1953) classic paper on open concepts. While, strictly speaking, I have to agree with my late friend Paul Feyerabend that all concepts are open, I agree with Lakatos' answer that "some of them are opener than others." How are we to reconcile the openness of theoretical concepts with the demand for severe tests? Answer: *Statisticize the openness*. Probability numbers will appear in the object language of our theory, which should not bother anybody familiar with classical psychometrics; probability numbers appear in all empirical sciences with which I am familiar.

My taxometric method was devised with this injunction in mind. We do not foolishly attempt a fake operational definition of schizophrenia, schizotypy, or schizotaxia in trying to test my theory. We would *like* to be able to identify with high confidence each individual in a family pedigree carrying the schizotaxic gene (on my conjecture, an autosomal dominant, completely penetrant for the neurological aberration, but only 10% penetrant for clinical schizophrenia). But we do not require this. What we *can* derive is theorems generated by a postulated model of a latent taxon, and a variety of internal consistency tests to tell us whether the model is correct and the inferred latent numerical values are acceptable within reasonable tolerances. I know of five large-scale projects on the psychophysiology, soft neurology, and mental status of schizophrenic probands and their families which will permit a definitive taxometric test of my schizotaxia theory. None of it will rely on refutations of H_0.

History of science shows that good convergence to theoretical entities, via their inferred numerical values, by multiple epistemic paths is more potent than pseudo-precise estimates via a single path. I would rather have four halfway decent avenues to a gene frequency than a maximum likelihood estimator via only one. If that shocks devout Fisherians, I point out that astronomy, physics, chemistry, and portions of geology and physiology were in an advanced scientific state before R. A. Fisher was born. I do not denigrate Fisher, who was a genius and a major contributor, but I complain of the Fisherian social scientists, who do not understand the difference between a Fisherian approach to the very special problems of agronomy—where there is negligible difference between the substantive theory of interest and the statistical hypothesis—and the general scientific method, which is not formulable in such narrow terms.

If my taxometric method were applied to a sample of 3000 patients that could be studied in a year or two by a batch of cooperative clinics (a methodological

practice which has become common in epidemiology), we would know with high confidence which *DSM* categories denote objectively existing taxa and which do not. I pick the value 3000 because my method needs a large N for us to be fairly confident of identifying any genuine taxon making up at least 1% of our clinical population; we can look for rarer conditions in subsequent investigation.

The failure of some in our profession to recognize solutions to quantitative problems or to apply them to controversies is partly due to self-selection in the social sciences for poor quantitative talent; but in psychology it is due even more to the lamentable lack of mathematical education of psychology majors. I had to learn the slight mathematics I know by taking 23 credits of college algebra, analytic geometry, differential and integral calculus, and probability theory. Today, almost all mathematics departments provide more condensed courses, which, for instance, minimize the details of conic sections in analytic geometry and have some social science examples in the calculus course instead of it all being about computing the volume of a football. But the pedagogical problem here involves a vicious circle, which I do not know how to cure. Students have lots of pressures and demands, they have to set priorities; and most humans are happier to take the easy way, which never means mathematics. A senior planning to go to graduate school contemplates his advisor, a tenured professor who never took college algebra, and does not know what a partial derivative is, or the inverse of a matrix. The student thinks, "Well, he seems to have done very well, and writes these articles and books without knowing any math, so why should I bother with it?" For obvious reasons, the advisor rarely tries to correct this pleasant, easy-going view. We never *find out* whether mathematically competent people in the soft areas like social psychology or psychopathology could invent stronger theories susceptible of severe tests, because, since there are not very many such strong theories (except in behavior genetics), it does not appear to the student that this would be a profitable undertaking.

Please do not misunderstand me. Unlike some *Psychometrika* authors who mainly display their virtuosity in manipulating the formalism, I am not pushing rigor in the mathematics and depth of mathematical understanding for its own sake. Even a PhD in theoretical physics is not normally expected to spin off proofs of theorems in pure mathematics that some derivations presuppose. But physicists are expected to understand the idea of continuity, and what a partial derivative is, and what is assumed about space when one applies the fundamental theorem of the integral calculus. Mathematical concepts that bear directly upon the interpretation of a concept in empirical science should be understood by the scientist. *Examples*: There would be some point in working through Pearson's derivation of the χ^2 probability integral, but not much, so I would not push it. But given our constant use of the normal curve, a psychology student should understand that the normal curve is what you get when you consider the expansion of $(p + q)^N$ for a large number of small slices, the basic notion of DeMoivre's famous theorem. It is sometimes a fine line, but the point is obvious.

Although quantitative reasoning is an *ability* (as contrasted with a scholastic achievement), when the factual situation becomes complicated, a person with high quantitative reasoning ability but who is woefully ignorant of mathematics cannot be expected to cope with it competently. To say breezily, "Well, I do not understand the math but I do understand the logic," is usually phony, because the "logic" of something like factor analysis or multidimensional scaling or my taxometrics *is* mathematical, and there is no cutesy way of getting around that. I once sat on a doctoral oral which involved factor analyzing data on children's personalities and parental child-rearing attitudes and practices. The psychometric instruments were devised by someone else, the interviews conducted by someone else, and the factor analysis done by our computer center. The student called them up and said he wanted some data factor analyzed. They asked him what solution to the rotation problem he wanted, and he asked them what was customary; they said, "Well, most people seem to like Kaiser's Varimax rotation," and he said, "Fine," and the next thing you know he has a PhD thesis. I am not complaining of file data—my own doctoral dissertation was based on MMPI file data, and sometimes that is unavoidable because you need a large *N* that would take 10 years to gather if you tested all these patients personally—but I did the statistics myself and I knew what I was doing. I asked the doctoral student what the Varimax rotation was, and he had no idea. He could not even tell me that the factors are uncorrelated. I asked him why there *is* a rotation problem in the first place, and he did not have a clue. All his dissertation demonstrated was that he could write literate English and could make a telephone call to the computer lab.

I am aware that this is a common state of affairs, but I say it is scandalous. It could never happen in a physics or chemistry department. It is not the student's fault; it is the fault of the advisor and of the whole nonmathematical intellectual tradition of soft psychology. But it is still wrong. After this egregious instance, I circulated a letter among the clinical advisors saying that from now on I was going to require not that students should do their own factor analysis with a desk calculator, but they should show rudimentary understanding of the mathematics, and they should know what certain mathematical operations on data sets could and could not prove about the real world, and why. (Result: I ceased to be put on most examining committees, an unintended consequence of which I did not complain.)

Yet, the vicious circle is perhaps curing itself by indirect means. Our clinical students discern that their ablest peers are quantitatively talented and mathematically informed, and that the same holds for faculty. As a result of self-selection by the alphas, I found some mathematical derivations of my co-author Niels Waller (PhD, Minnesota 1990) hard-going, and I am second author of our recent book on taxometric analysis (Waller & Meehl, 1998) because the multivariate generalization is his. But this encouraging process would accelerate if the faculty would forget their own insecurities and stiffen up mathematical requirements for the PhD in psychology. In the methodology seminar Will Grove and I teach, we find that about ¼ of the graduate students have had a course in logic,

and nearly all have had calculus. The proportion who have studied logic is about as it was when I began graduate work 57 years ago, but the calculus percentage has greatly increased. I find this encouraging.

I conclude with a quote from one of the greats, Edward Lee Thorndike, which, while not literally true, is a safer illusion than the opposite one found in the non-quantitative subculture. "Our ideals may be as lofty and as subtle as you please, but if they are real ideals, they are ideals for achieving something; and if anything real is ever achieved, it can be measured. Not perhaps now, and not perhaps fifty years from now; but if a thing exists, it exists in some amount; and if it exists in some amount, it can be measured."

REFERENCES FOR CHAPTER 18, QUANTITATIVE THINKING

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437. Reprinted in D.E. Morrison & R.E. Henkel (Eds.), *The significance test controversy* (pp. 231-251). Chicago, IL: Aldine, 1970.

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, 134, 596-602.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Hogben, L. (1968). *Statistical theory*. New York: Norton. (Original publication 1957)

Jöreskog, K.G., & Sörbom, D. (2001). LISREL 8.50 *user's reference guide* [Computer software manual]. Chicago, IL: Scientific Software.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-159. Reprinted in D.E. Morrison & R.E. Henkel (Eds.). *The significance test controversy* (pp. 267-279), Chicago, IL: Aldine, 1970.

Meehl, P.E. (1954/1996). *Clinical versus statistical prediction: a theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press. Reprinted with new Preface, 1996, by Jason Aronson, Northvale, NJ.

Meehl, P.E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115. Reprinted in D. E. Morrison & R. E. Henkel (Eds.), *The significance test controversy* (pp. 252-266). Chicago, IL: Aldine, 1970.

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.

Merton, R.K. (1973). *The sociology of science*. Chicago, IL: Univer. of Chicago Press.

Pap, A. (1953). Reduction-sentences and open concepts. *Methodos*, *5*, 3-30.

Pearson, K. (1895b). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, Series V.1, 157-175.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416-428. Reprinted in D.E. Morrison & R.E. Henkel (Eds.), *The significance test controversy* (pp. 216-230). Chicago, IL: Aldine., 1970.

Serlin, R.C., & Lapsley, D.K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist*, *40*, 73-83.

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

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

Wilkinson, L., & Task Force on Statistical Inference APA Science Directorate (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

pdf by LJY, September 2010