Brit. J. Phil. Sci. **50** (1999), 283-295

#173

173.

How to Weight Scientists' Probabilities Is Not a Big Problem: Comment on Barnes

Paul E. Meehl

ABSTRACT

Assuming it rational to treat other persons' probabilities as epistemically significant, how shall their judgements be weighted (Barnes [1998])? Several plausible methods exist, but theorems in classical psychometrics greatly reduce the importance of the problem. If scientists' judgements tend to be positively correlated, the difference between two randomly weighted composites shrinks as the number of judges rises. Since, for reasons such as representative coverage, minimizing bias, and avoiding elitism, we would rarely employ small numbers of judges (e.g. less than 10), the difference between two weighting systems becomes negligible. Suggestions are made for quantifying verisimilitude, identifying 'types' of scientists or theories (taxometrics), inferring latent factors, and estimating reliability of pooled judgements.

A challenging article by Eric Christian Barnes ([1998]) argues for the rationality of treating other persons' probabilities as epistemically significant. I do not address the qualitative merits of this position, about which I have no settled opinion.[1] My purpose here is to offer a psychologist's reassurance to Barnes in respect to one of his major concerns: how to weight the opinions of a group of scientists who will surely disagree in their subjective probabilities.

If we do not give equal weight to the subjective probabilities of Scientists X_1, X_2,…X_n, we would wish to weight them differentially on the basis of their several 'competencies'. Competence being itself a complex notion, Barnes considers two ways of judging it, which, I gather, he does not claim to be exhaustive. One way ('mapmakers') involves a fairly complex indexing of the judges' knowledge bases. The other ('probabilistic jury') relies on their

[1] It is, however, obvious that one situation routinely permits such head-counting as rational and unavoidable, namely, determination of what is to be asserted in textbooks. To a considerable extent research funding is another. By indirection, hiring and promoting faculty is a third. Trusting referees of journal submissions is a fourth, currently under critical scrutiny.

different histories of success, with the attendant criterion–circularity problem. He then discusses mixing kinds of weights and some pitfalls of the method.

Barnes quite properly implies the need for employing a sufficient number of such scientist-judges, and this guideline is the basis for my reassuring thesis. It is an accepted, non-controversial principle of classical psychometrics that *when a sizeable number* (*say*, $n > 10$) *of variables are employed in a linear equation to get a composite score, the weights do not matter appreciably* (Bloch and Moses [1988]; Burt [1950]; Dawes [1979, 1988], Ch. 10; Dawes and Corrigan [1974]; Einhorn and Hogarth [1975]; Gulliksen [1950]; Laughlin [1978]; Richardson [1941]; Tukey [1948]; Wainer [1976, 1978]; Wilks [1938]). The cliché familiar to psychometricians is in the homespun title of a classic paper by Wainer ([1976]), 'Estimating Coefficients in Linear Models: It Don't Make No Nevermind.' Analytic, Monte Carlo,[2] and real data studies over some sixty years have uniformly supported this strong generalization, so that psychometricians today routinely rely on it when building psychological tests or forming test batteries.

Although Barnes does not specify a minimum number of scientists for weighted pooling, his phrase 'epistemic pluralism' implies more than one. Reflecting on the idea's rationale leads me to opine that (1) no definite number can reasonably be set; (2) *ceteris paribus*, the more the better; (3) it would be inappropriate in writing a textbook to rely on only one or very few judges; (4) yet it may be rational in some circumstances to rely on a single judge if that is all you have (e.g. on a disputed point in quantum mechanics I would rely on Dirac's judgement rather than on the pooled judgements of ten psychologists); (5) one reason for collecting multiple judgements is the hope of lessening bias factors operating in individuals. Barnes agrees with these statements (personal e-mail communications, October 1998).

Given that it is neither rationally possible nor epistemically necessary to fix a minimum number as appropriate for the procedure, we can nevertheless say some useful things about rough ranges. (We do not succumb to the slippery slope argument, which in this situation would be fallacious.) The basic motivation behind pooling judgements in this context is that different scientists have partly non-overlapping information (the map analogy), but they also differ in biases, cleverness, 'empirical intuition', conscientiousness in the rating task, fear of being mistaken, theoretical optimism-pessimism, and so on. We hope that by collecting numerous opinions we will, so to say, 'cover the waterfront', that the numerous non-rational factors impelling theory appraisals to and fro will tend, statistically speaking, to cancel out. The causal–statistical model here is shooting at a target, where such factors as wind, motor steadiness,

[2] The Monte Carlo method puts 'artificial data' into the computer's memory and samples randomly from the hypothetical population to answer quantitative questions that are, or seem, intractable to analytic solution.

visual acuity, and non-optimal sight adjustment displace shots from the optimum, but the centre of gravity of the bullet holes is in the bulls-eye. It is the empirical fact of inter-scientist *differences* that even suggests the idea of handling epistemic pluralism by a weighted sum of judgements. (This has a short, easy, ineluctable proof: if scientists did not differ, the judgement of any one scientist tells the whole story, as they all judge the same.)

Our situation is similar to that of the public opinion pollster, who stratifies the sample with respect to those demographic variables known to influence political opinion (e.g. income, education, sex, age, race, religion, geography). This consideration helps us in choosing a 'reasonable' number of judges with the aim of reducing bias and unrepresentativeness. Suppose there are four characteristics of scientists in a given field that are known, or plausibly feared, to operate as strong biasing influences in scientific judgement about a class of issues. For example, in psychopathology the views of clinicians about aetiology and treatment of mental disorder are strongly clustered as 'biotropes' (emphasizing genes, organic disease, and drug therapy) versus 'sociotropes' (emphasizing psychodynamics and childhood family influences). This pervasive orientation dimension, that affects opinion on a variety of issues, has been shown to correlate with age, profession (MD or Ph.D.), geography (Midwest USA versus either coast) and ethnicity (Gentile/Jewish). As in Fisher's ([1971]) experimental design in agronomy and medicine, we wish to avoid the statistician's *confounding*, where pairs of factors always appear together and the statistics cannot tease their influences apart. Thus, if there are dichotomous values of these four demographic variables, we prefer to have each dichotomous value paired with all combinations of the others, requiring $2^4 = 16$ configurations ('cells', in agronomy, each containing several individuals). If we accepted a minimum representation of one scientist per configuration, we would employ 16 scientists, on 'waterfront coverage' grounds.

The assumptions and formulas derived in the above list of weighting references vary somewhat, the basic assumption being the very weak one of 'positive manifold'. That is, the predictor variables are positively correlated because different judges, varying in their competence, data bases, and biases are nevertheless roughly aiming at the same thing (verisimilitude—or, if instrumentalists, long-term survival). Inasmuch as the whole business is statistical, a few small negative correlations departing from perfect positive manifold does not vitiate the theorem. That pairs of competent scientists' evaluations over theory sets would rarely correlate strongly negative seems a safe working assumption. If that assumption were untrue, we would quickly detect it in applying Barnes' approach, and the rationale for the approach would be put in grave doubt.

According to psychometrician Cronbach, whose opinion is heavily weighted in my profession, the best developments are by Burt ([1950]) and Richardson ([1941]).

Here I will use Burt's because it is somewhat stronger and formulated in terms of tests rather than items. Suppose we have a system of predictor variables x_1, x_2,…x_n combined using weights β_1, β_2,… β_n yielding a composite index $\tilde{x}$. This composite might be 'optimal' for predicting an external variable y (as in personnel selection) or for inferring the values of a latent theoretical factor (such as g [= general intelligence], inferred from the dozen subtest scores of an IQ test). But it may be optimal for neither, as in Barnes' situation. Let us now assign a different set of β-weights randomly to the xs and compute a new (scrambled, degraded, nonoptimal) composite $\tilde{x}'$ from our fixed set of numerical xs. This new $\tilde{x}'$ will of course not agree perfectly with the original $\tilde{x}$. The expected value of the correlation[3] between the two composites is

$$\bar{r}_{\tilde{x}\tilde{x}'} = 1 - \frac{(1-\bar{r})}{2n\bar{r}}\left(\frac{\sigma_{w_1}^2}{\bar{W}_1^2} + \frac{\sigma_{w_2}^2}{\bar{W}_2^2}\right)$$

where $\bar{r}$ = mean pairwise correlation of the variables, n = number of variables in the composite, σ_{w_1} and σ_{w_2} = standard deviations of the weights in the two weighting systems, $\bar{W}_1$ and $\bar{W}_2$ = the mean weights in the two systems.[4] McCormack ([1956]) warns that for composites of *very* high validity, the weights begin to matter appreciably; so if it should (surprisingly?) turn out that *pooled* scientists correlate $R = .95$ with verisimilitude, Barnes' worry returns.[5]

Numerical examples for various parametric situations can be found in the references cited above and in Meehl ([1990], [1992a]), where there is extended discussion of this topic. A handy mnemonic rule of thumb is that the expected correlation between two composites of n variables with randomly assigned weights is $\bar{r}_{w_1 w_2} \cong 1 - 1/n$. The variance (squared standard deviation) is approximately $1/n^2$.

[3] The correlation coefficient measures how closely two quantitative variables covary. It ranges from $r = -1$ (perfect negative relation) through $r = 0$ (no relation) to $r = +1$ (perfect positive relation). If each variable is expressed in standard score form, as deviation from its own mean divided by its standard deviation,

$$x_i = \frac{X_i - \bar{X}}{SD_X} \qquad y_i = \frac{Y_i - \bar{Y}}{SD_Y}$$

then r is the slope of the best-fitting straight line predicting Y from X. Representing variables by vectors in a hyperspace, the correlation between any pair of variables is the scalar product of their vectors, so in positive manifold the cosines of the central angles are all < 1.

[4] This is the correct formula, given incorrectly in Meehl ([1992a], p. 456). The standard deviation (σ, or SD) of a distribution of numbers X_1, X_2,…X_n is the root mean square of their deviations from their mean $\bar{X}$,

$$SD = \sqrt{\frac{1}{N}\sum^{N}\left(X_i - \bar{X}\right)^2} \ .$$

The usual statistician's measure of variability ('scatter', dispersion), it is analogous to a flywheel's radius of gyration in mechanics.

[5] Pragmatically, the proposed epistemic pooling would occur importantly in the early or middle life of a theory, where disagreements are considerable. We hardly need to compute a statistical composite of physicists' opinions about GTR, or psychologists' views of phrenology, today.

For almost all empirical parametric situations this rough approximation tends to underestimate the agreement, thus for 10 judges $\bar{r} > .90$, for 20 judges $\bar{r} > .95$, and for 50 judges $\bar{r} > .98$.

Using Burt's more accurate formula, suppose we have judgements by ten scientists whose opinions have weights (however assigned)

$$w_i = .05, .15, .20, .25, .25, .25, .25, .30, .35, .40,$$

a considerable but not unreasonable dispersion of their estimated 'competencies'. The positive manifold condition tends to hold down this dispersion, and for most real data in social science the standard deviation of weights is less than half their mean.[6] For measures of achievement, ability, and personality one expects dispersions 15%–25% of the mean. Let us reassign these 10 weights randomly to the 10 scientists. Then the expected value of the correlation between the optimal and degraded composites is $r_{ww'} = .96$, nearly perfect agreement—scrambling the wide-ranging weights has negligible effect on the resultant ordering of pooled judgements. For a wide variety of equally reassuring numerical examples, see Meehl ([1992a], Appendix 1, pp. 459–62). Even if the average pairwise judge correlation were as improbably low as $\bar{r} = .10$, the expected agreement between the two weighting systems is still reassuringly high at $r = .87$. Given the subjective and stochastic character of the problem, fretting about small departures from perfect agreement would be 'cutting butter with a razor', as my logical positivist mentor Herbert Feigl used to say in warning social scientists against it.[7]

It is tempting to assume that the weighting problem can be avoided by merely adding the component judgements, but this is incorrect. Suppose n scientists judge a set of theories by rating each one on a seven-point rating scale (Guilford [1954], Ch. 11). If we simply add these 'raw score' ratings to get a composite judgement, the *nominal* (unit) *weights* (chosen by us *a priori*, as if on some variant of the Principle of Insufficient Reason) will never be identical with the *effective weights*,[8] as the latter are now determined willy-nilly, not by us but by the facts. Theorems of classical psychometrics show that, defining a variable's 'effective weight' in a composite in terms of the proportion of the composite's variance that the variable predicts (its shared variance with the composite), the effective weights depend in a rather complicated way on their several dispersions (standard deviations) and their pairwise intercorrelations. A scientist whose ratings are spread widely over the rating scale
will receive a heavier weight, and one whose judgements agree more with the other raters

[6] If, as customary in social science, the predictand or the composite is standardized with mean = 0 and standard deviation = 1, the sum of squared weights cannot exceed 1, putting a heavy lid on their dispersion. Human judgements belong to social science, and I have no reason to think that this situation is markedly unlike the thousands that have been studied.

[7] I have been told the current phraseology among physicists is 'Cutting SPAM with a laser'.

[8] 'Effective weights' does not mean *efficient* weights, those that best predict a Gold Standard Criterion—in this situation, objective truth (or verisimilitude), not directly accessible to us. For an accessible surrogate see Meehl ([1992a]).

will receive a heavier weight. The algebra of multiple correlation operates on the data to generate the effective weights. If we express each rater's judgements in standard score form with mean = 0 and standard deviation = 1.00, the various judges' ratings are now partially 'equated' in the sense of having an identical metric uninfluenced by the difference in their dispersions.[9] But the effective weights are still unequal because of differences in the inter-rater correlations. When variables in standard scores are assigned 'equal' (= standardized) nominal weights in a composite, each variable contributes to the total variance in proportion to its correlation with the composite (Richardson [1941], p. 384).

Giving epistemic relevance to scientists' judgements presupposes (conjectures?) that such judgements are correlated with verisimilitude and/or long-term instrumental success.[10] In classical psychometrics we classify weighting systems as *a priori* (e.g. value, intuitive, theoretically plausible, rational, commonsensical), as statistically predictive of an external criterion (e.g. academic success, income, parole violation, response to psychotherapy), or as internal statistical structure (e.g. factor analysis, taxometrics). Barnes' two weighting methods are roughly of the first and second kinds. In the second method we can correlate scientists' judgements with theories' long term success (e.g. consensus ensconcement versus discard for fifty years; see Faust and Meehl [1992] and Meehl [1992a, 1998]). In naturalized epistemology we may treat the dichotomy of fifty-year ensconcement/discard as proxy for Peirce's 'truth' criterion of ultimate acceptance by those who investigate. A few long-term reversals of half-century consensus is harmless, because the entire weighting procedure is explicitly statistical [= single case fallible] *in principle* (for elaboration see Meehl [1992a, 1998]). For naturalized epistemologists this

[9] The standard score of judge *j*'s rating X_{ji} of theory *i* is

$$\frac{X_{ji} - \overline{X}_{ji}}{SD_{ji}}$$

where SD_{ji} is the standard deviation of that judge's ratings over the class of *n* theories *i* = 1, 2,…*n*. This standardizing transformation does not assume a Gaussian ('normal') distribution, although some familiar probabilities inferred from it do involve integrals of the Gaussian density function.

[10] The meta-concept *verisimilitude* is indispensable in all scientific thinking, as it is in journalism, business, law courts, and common life. Our current inability to provide a rigorous explication cannot justify discarding the explicandum; this persistent problem simply invites the logician to keep working on it (e.g. Niiniluoto [1998]), as on other deep, hard, important concepts. Philosophers regularly write about induction, proof, disproof, confirmation, falsification, implicit definition, the causal nexus, mental events, intentionality, reference, reduction, probability, nonconstructive proofs regarding transfinite cardinals, set theory, formalism versus logicism in mathematics, Platonism regarding numbers, dispositions, possible worlds, bridge laws, operational definitions, psychological causality, determinism, locality and entanglement in quantum mechanics, emergence, supervenience, *not one of which* has been rigorously explicated to everybody's satisfaction. It is puzzling why many philosophers treat verisimilitude so harshly and cavalierly compared with these other concepts and conjectures as explicanda. As an ideal 'criterion,' either a realist or instrumentalist framework will do in appraising a weighting procedure. The closeness of their agreement becomes an interesting theoretical and empirical question. I do not address fictionism because I do not understand that position.

seeming circularity (on Feyerabend's 'big circle') is of course not a vicious circularity.

If, despite the negligible effect of weights under realized scientific conditions, we desire to weight differentially but do not trust either an a priori or external criterion method, it can still be done by the internal relations approach. In the classic paper by Wilks ([1938]) that gives $(1 - 1/n)$ as the expected correlation between composites using two weighting systems, a procedure is derived that optimizes the *stability* (psychometric 'reliability') of composite orderings. In the present situation this amounts to inferring the most trustworthy ordering of the theories being judged, conjecturing that the scientists' judgements are statistically 'valid' for that unknown, underlying dimension of verisimilitude. In the simplest (ideal) case, one large latent statistical factor would pervade the judgements, the remainder of the variance being random or contributed by very few factors of negligible contribution. Wilks shows that this corresponds to the *principle components* procedure of Hotelling ([1933]) if we weight the fallible indicator variables on the basis of their projections on the first principal component. If the first principal component leaves considerable rating correlations unaccounted for, the statistical procedure *multiple factor analysis* may be in order. That procedure aims to discern latent mathematical factors underlying a matrix of correlations, estimating how many such factors are needed to account for most of the reliable covariance among observed indicators, and it infers the *factor loadings* of each manifest indicator on the several factors. If, as seems likely, different scientists pay quite different attention to the various properties and relations of theories, a hierarchical factor analysis would be preferable, in which the presumed big general second-order factor (verisimilitude) pervades the judgements 'via' several correlated first-order factors that correspond to scientists' different epistemic emphases.[11]

[11] Jensen and Weng ([1994]) have shown that the several procedures for hierarchical factor analysis yield almost identical results, surely more accurate than the present problem demands. The *locus classicus* of factor analysis is Louis Leon Thurstone, *The Vectors of Mind* ([1935]), revised as *Multiple Factor Analysis* ([1947]). He provides a preliminary chapter on the necessary theory of vectors and matrices, but for non-mathematical readers it can be a bit difficult despite his expository clarity. A highly regarded but difficult standard treatise is Harman ([1960]). Good introductory treatments are Gorsuch ([1983]), most often suggested by my psychometric colleagues); Reyment and Jöreskog ([1993]); Goldberg and Digman ([1994]); Thomson ([1951]); Long ([1988]); and McDonald ([1985]). The most generally 'philosophical' treatment is Burt ([1940]). A good exposition that does not rely on matrix algebra or hyperspace representation of factors is Peters and Van Voorhis ([1940]). A short, clear introduction is Adcock ([1954]), but some of the mathematical material is out-dated because it is pre-computer, and Thurstone's centroid method has been abandoned. Philosophers should not be put off by the references to 'tests', 'abilities', 'intelligence', 'subtests' which occur so frequently because factor analysis was invented by psychologists. The mathematics is quite general, and may be useful whenever one studies fallible indicators whose stochastic relations arise from unknown latent quantitative influences. In the present instance the 'subjects' are scientific theories, their 'factor scores' are their objective properties—especially their verisimilitude—the fallible 'test scores' are scientists' judgements, and the several 'tests' (indicators) are the several scientists.

An interesting possibility arises if we tentatively accept a survival criterion having ascertained, without using it, the scientists' weights in terms of the first principal component. If, as Barnes hopes, scientist judgements have epistemic validity, then the weighted composite judgement should correlate significantly higher with the ensconcement criterion than does a unit standardized or *a priori* weighting. We have here a sort of metatheoretical 'inference to the best explanation', the best explanation being that the scientists' judgements correlate as they do because they all correlate, albeit not equally, with verisimilitude.

In treating theory verisimilitude as a latent 'factor' that 'loads' (influences, affects, controls) a batch of human ratings, am I illicitly reifying statistical factors? The ontological status of statistical factors is an obscure and disputed question which I cannot properly examine here.[12] Shortly and dogmatically put, some factors are clearly fictional (like magnetic lines of force or suicide waves), others equally clearly real (like blood glucose level), and—in social science—many of unclear status (like social class, or the general intelligence factor *g*, whose heritable component *could* be literally the physical count of 'bright-generating' polygenes, but not if one is a strong anti-hereditarian about intelligence).

Verisimilitude, however, is an ontological metapredicate, by definition. What of the objection that we are treating *reasons* as *causes*, which some say they cannot be? (I cannot discuss Popper's World 3, which I do not comprehend, although I admit he is there addressing one of the great philosophical problems.) Short answer, I hope adequate for Barnes' purpose: In identifying verisimilitude with a statistical factor underlying scientists' judgements about theories, one speaks loosely. What are causally efficacious in the biological system 'scientific community in the world' are the objective events participated in by the entities about which the theory speaks. These events initiate causal chains that impinge on scientists' sensory receptors, causing events in scientists' brains.[13] Scientists' brain-events result in overt behaviours (speaking, writing) that influence other scientists' brains, so they speak, and write, and set up apparatus, thereby eliciting further novel causal chains from the theory-described entities, and

[12] But see Meehl ([1993]); Burt ([1940]); Cronbach and Meehl ([1955]); Thomson ([1951]). Factor analysts disagree about this, and often avoid it with superficial pseudo-solutions. Many psychologists are inconsistent, espousing a vulgar fictionism in their meta-talk but a poorly formulated quasi-realism in their object-language discourse. For example, one cannot easily rationalize engaging in disputes about how to rotate the factor axes (e.g. psychological interpretation? parsimony? what kind?) if a factor matrix is *nothing but* an 'economical representation' of the original correlation matrix. There's nothing economical (except page space) or 'conveniently summarizing' about communicating a matrix which must be matrix-multiplied by its transpose to obtain a mere approximation of the observational matrix it purports to 'represent' (Meehl [1954], pp. 12–14).

[13] I here adopt monistic language as the coin of the realm, but nothing hinges on this. An event-dualist (or even a substance-dualist, see Meehl [1966, 1989]) may substitute *brain-mind* for *brain*, so long as causal influence, be it nomological or stochastic, is allowed on and by that complex entity.

so it goes. Whatever the correct metaphysics of reasons (logical relations) may be,[14] it is obvious that the *tokening* of sentences *expressing* propositions that *constitute* reasons—less precisely put, the speaking of reasons, thinking of reasons, writing of reasons, hearing reasons, reading reasons—are events in the World 2 of us cognizing animals. One need not be an eliminative materialist to understand that truth, as we tremble in our boots because the mushroom cloud is the terminus of a causal chain originating in Einstein's head when he wrote $E = mc^2$. Any ordinary language analysis that purports to prove this kind of sequence is impossible has to be defective.[15]

Summarizing, in conceptualizing verisimilitude as a factor that loads scientists' ratings of theories, we must unpack the psychometrician's claim thus: Theories of high verisimilitude tend statistically to have properties and relations that are 'better' in scientists' eyes than low verisimilitude theories; hence they tend to receive more favourable ratings. If some scientists are better at the judging task than others, this tendency will be reflected in their receiving heavier weight, whether the weighting is based on an external criterion or, lacking a criterion, on the group judgements' internal statistical structure. But this is likely to be a psychometric nicety because in almost all realized judgement situations where a considerable number of scientists are involved, the weights do not matter enough to be worth bothering with.

Philosophers and historians of science may wonder whether these psychometric procedures, or Barnes' basic idea of a linear pooling of judgements, can do justice to a situation where there are marked *qualitative* differences among sub-groups of scientists (e.g. daring deductivists versus safe-playing inductivists; reductionists versus anti-reductionists in biological and social sciences). The theorems described are algebraic identities and do not hinge on denying such 'typological' possibilities. If this is of intrinsic interest, appropriate methods exist for determining whether types (taxa, syndromes, species, 'natural kinds', *non-arbitrary categories*) exist. For taxometrics, see Meehl ([1995], [1992b]) and Meehl and Golden ([1982]). It can be shown that taxometric analysis is a special case of factor analysis (Waller and Meehl [1998]).[16] Another approach to quantifying a typology is cluster analysis (Sneath and Sokal [1973]; Hartigan [1975]; Arabie, Hubert, and De Soete [1996]).

We may properly view the judgements of different scientists as aiming, more or less well, at the same epistemic target, although some are perhaps doing it better than others. If disagreements are treated psychometrically as *unreliability*, pooling the

[14] In what realm of being would Goldbach's Conjecture be false, if it is, assuming we never find the number that falsifies it nor a general proof of its falsity? God forbid that I, a non-logician, should try to discuss that monster puzzle here.

[15] On the related problem, posed by Popper, that mind-brain determinism precludes rationality, see Feigl and Meehl ([1974]) and Meehl ([1970], [1989]).

[16] Some psychologists assume that the conventional factor analytic procedures can answer the question of taxonicity, but this is a mathematical mistake.

judgements is a way of boosting reliability. If we have the judgements of only a few scientists (rating a batch of theories or single experiments), we can estimate the reliability of a larger pooled judgement via the *Spearman–Brown Prophecy Formula*,

$$r_{\mathrm{II}} = \frac{kr_{\mathrm{I}}}{1+(k-1)r_{\mathrm{I}}}$$

where r_{I} is single judgement reliability and k is the number of judges to be pooled. This formula, invented (Brown [1910]; Spearman [1910]) to predict the boosted reliability of a lengthened mental test, has turned out to be quite accurate when the elements are not test items but human judgements.[17] For example, if the reliability of a schoolteacher's rating of children's intelligence is $r = .60$, we need to pool judgements of seven teachers to equal the reliability of an IQ test ($r \cong .90$).

If I may generalize from the instigating issue of how to weight different scientists' subjective probabilities, perhaps the larger lesson here is that the psychometrician can sometimes play a useful role as *ancilla philosophiae*.

University of Minnesota
Department of Psychology, N218 Elliott Hall
75 East River Road
Minneapolis, MN 55455–0344
USA
pemeehl@umn.edu

[17] A rather trivial result, since any physical entities satisfying the axioms of psychometric theory must satisfy the theorems. The necessary and sufficient condition is statistical homogeneity of the old and new items, i.e., that they have a similar distribution of item difficulties and intercorrelations.

References

Adcock, C. J. [1954]: *Factorial Analysis for Non-mathematicians*, Carlton, Victoria, Australia: Melbourne University Press.

Arabie, P., Hubert, L. J., and De Soete, G. (*eds*) [1996]: *Clustering and Classification*, River Edge, NJ: World Scientific.

Barnes, E. C. [1998]: 'Probabilities and Epistemic Pluralism', *British Journal for the Philosophy of Science*, **49**, pp. 31–47.

Bloch, D. A., and Moses, L. E. [1988]: 'Nonoptimally Weighted Least Squares', *American Statistician*, **42**, pp. 50–53.

Brown, W. [1910]: 'Some Experimental Results in the Correlation of Mental Abilities', *British Journal of Psychology*, **3**, pp. 296–322.

Burt, C. [1940]: *The Factors of the Mind: An Introduction to Factor-Analysis in Psychology*, London: University of London Press.

Burt, C. [1950]: 'The Influence of Differential Weighting', *British Journal of Psychology, Statistical Section*, **3**, pp. 105–23.

Cronbach, L. J. and Meehl, P. E. [1955]: 'Construct Validity in Psychological Tests', *Psychological Bulletin*, **52**, pp. 281–302. Reprinted in H. Feigl and M. Scriven (*eds*) [1973], *Minnesota Studies in the Philosophy of Science, Vol. 1: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, Minneapolis, MN: University of Minnesota Press, pp. 174–204.

Dawes, R. M. [1979]: 'The Robust Beauty of Improper Linear Models in Decision Making', *American Psychologist*, **34**, pp. 571–82.

Dawes, R. M. [1988]: *Rational Choice in an Uncertain World*, Chicago, IL: Harcourt Brace Jovanovich.

Dawes, R. M., and Corrigan, B. [1974]: 'Linear Models in Decision Making', *Psychological Bulletin*, **81**, pp. 95–106.

Einhorn, H. J., and Hogarth, R. M. [1975]: 'Unit Weighting Schemes for Decision Making', *Organizational Behavior and Human Performance*, **13**, pp. 171–92.

Faust, D., and Meehl, P. E. [1992]: 'Using Scientific Methods to Resolve Enduring Questions within the History and Philosophy of Science: Some Illustrations', *Behavior Therapy*, **23**, pp. 195–211.

Feigl, H. and Meehl, P. E. [1974]: 'The Determinism-Freedom and Mind-Body Problems', in P. A. Schilpp (*ed.*), *The Philosophy of Karl Popper*, LaSalle, IL: Open Court, pp. 520–59.

Fisher, R. A. [1971]: *The Design of Experiments*, 9th edn, New York: Hafner.

Goldberg, L. R., and Digman, J. M. [1994]: 'Revealing Structure in the Data: Principles of Exploratory Factor Analysis', in S. Strack and M. Lorr (*eds*), *Differentiating Normal and Abnormal Personality*, New York: Springer, pp. 216–42.

Gorsuch, R. L. [1983]: *Factor Analysis*, 2nd edn, Hillsdale, NJ: Erlbaum.

Guilford, J. P. [1954]: *Psychometric methods*, New York: McGraw-Hill.

Gulliksen, H. [1950]: *Theory of Mental Tests*, New York: Wiley and Sons.

Harman, H. H. [1960]: *Modern Factor Analysis*, Chicago, IL: University of Chicago Press.

Hartigan, J. A. [1975]: *Clustering Algorithms*, New York: Wiley.

Hotelling, H. [1933]: 'Analysis of a Complex of Statistical Variables into Principal Components', *Journal of Educational Psychology*, **24**, pp. 417–41, 498–520.

Jensen, A. R., and Weng, L. [1994]: 'What Is a Good *g*?', *Intelligence*, **18**, pp. 231–58.

Laughlin, J. E. [1978]: 'Comment on "Estimating Coefficients in Linear Models: It Don't Make No Nevermind"', *Psychological Bulletin*, **85**, pp. 247–53.

Long, S. J. [1988]: *Confirmatory Factor Analysis: A Preface to LISREL*, Beverly Hills, CA: Sage.

McCormack, R. L. [1956]: 'A Criticism of Studies Comparing Item-Weighting Methods', *Journal of Applied Psychology*, **40**, pp. 343–4.

McDonald, R. P. [1985]: *Factor Analysis and Related Methods*, Hillsdale, NJ: Erlbaum.

Meehl, P. E. [1954]: *Clinica Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, Minneapolis, MN: University of Minnesota Press; reprinted with new Preface [1996], Northvale, NJ: Jason Aronson.

Meehl, P. E. [1966]: 'The Compleat Autocerebroscopist: A Thought-Experiment on Professor Feigl's Mind-Body Identity Thesis', in P. K. Feyerabend and G. Maxwell (*eds*), *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*, Minneapolis, MN: University of Minnesota Press, pp. 103–80.

Meehl, P. E. [1970]: 'Psychological Determinism and Human Rationality: A Psychologist's Reactions to Professor Karl Popper's "Of Clouds and Clocks"', in M. Radner and S. Winokur (*eds*), *Minnesota Studies in the Philosophy of Science: Vol. IV. Analyses of Theories and Methods of Physics and Psychology*, Minneapolis, MN: University of Minnesota Press, pp. 310–72.

Meehl, P. E. [1989]: 'Psychological Determinism or Chance: Configural Cerebral Autoselection as a Tertium Quid', in M. L. Maxwell and C. W. Savage (*eds*), *Science, Mind, and Psychology: Essays in Honor of Grover Maxwell*, Lanham, MD: University Press of America, pp. 211–55.

Meehl, P. E. [1990]: 'Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant Using It', *Psychological Inquiry*, **1**, pp. 108–41, 173–80.

Meehl, P. E. [1992a]: 'Cliometric Metatheory: The Actuarial Approach to Empirical, History-Based Philosophy of Science', *Psychological Reports*, **71**, pp. 339–467.

Meehl, P. E. [1992b]: 'Factors and Taxa, Traits and Types, Differences of Degree and Differences in Kind', *Journal of Personality*, **60**, pp. 117–74.

Meehl, P. E. [1993]: 'Four Queries about Factor Reality', *History and Philosophy of Psychology Bulletin*, **5**(No. 2), pp. 4–5.

Meehl, P. E. [1995]: 'Bootstraps Taxometrics: Solving the Classification Problem in Psychopathology', *American Psychologist*, **50**, pp. 266–75.

Meehl, P. E. [1998, in preparation]: Rational Realist Reliabilism: Probabilifying Theories as a Side-Benefit of Renaturalizing Epistemology.

Meehl, P. E., and Golden, R. [1982]: 'Taxometric Methods', in P. Kendall and J. Butcher (*eds*), *Handbook of Research Methods in Clinical Psychology*, New York: Wiley, pp. 127–81.

Niiniluoto, I. [1998]: 'Verisimilitude: The Third Period', *British Journal for the Philosophy of Science*, **49**, pp. 1–29.

Peters, C. C. and Van Voorhis, W. R. [1940]: *Statistical Procedures and Their Mathematical Bases*, New York: McGraw-Hill.

Reyment, R. A., and Jöreskog, K. G. [1993]: *Applied Factor Analysis in the Natural Sciences* (2d *ed*.), New York: Cambridge University Press.

Richardson, M. W. [1941]: 'The Combination of Measures', in P. Horst, *Prediction of Personal Adjustment*, New York: Social Sciences Research Council, Bulletin No. 48, pp. 379–401.

Sneath, P. H. A. and Sokal, R. R. [1973]: *Numerical Taxonomy*, San Francisco, CA: Freeman.

Spearman, C. [1910]: 'Correlation Calculated with Faulty Data', *British Journal of Psychology*, **3**, pp. 271–95.

Thomson, G. [1951]: *The Factorial Analysis of Human Ability*, 5th edn., London: University of London Press.

Thurstone, L. L. [1935]: *The Vectors of Mind*, Chicago, IL: University of Chicago Press.

Thurstone, L. L. [1947]: *Multiple Factor Analysis*, Chicago, IL: University of Chicago Press.

Tukey, J. W. [1948]: 'Approximate Weights', *Annals of Mathematical Statistics*, **19**, pp. 91–2.

Wainer, H. [1976]: 'Estimating Coefficients in Linear Models: It Don't Make No Nevermind', *Psychological Bulletin*, **83**, pp. 213–17.

Wainer, H. [1978]: 'On the Sensitivity of Regression and Regressors', *Psychological Bulletin*, **85**, pp. 267–73.

Waller, N. G. and Meehl, P. E. [1998]: Multivariate Taxometric Procedures: Distinguishing Types from Continua, Newbury Park, CA: Sage.

Wilks, S. S. [1938]: 'Weighting Systems for Linear Functions of Correlated Variables When There Is No Dependent Variable', *Psychometrika*, **3**, pp. 23–40.

pdf by ljy August 2002