JOURNAL OF PERSONALITY ASSESSMENT, *85*(2), 134–140

What If There Were No Psychometrics?: Constructs, Complexity, and Measurement

Paul Barrett

Department of Management and Employment Relations
University of Auckland

McGrath (2005/this issue) has published a very significant article; it is significant because it examines the substantive issue of construct validity in a simple and informative way, noting the confusions apparent in how constructs and their variables are defined, used, and interpreted within the domain of personality and individual differences. This is a careful, thoughtful, article that quietly proceeds to outline the problems with current thinking and approaches to defining and using constructs in psychology. McGrath also suggests how some of these problems might be addressed. I note that McGrath appears to miss what might be considered an obvious rejoinder to some of his arguments, that is, the use of latent variable and item response theory. Some consideration is given to these arguments. However, a cursory examination shows that although these new methodologies offer many opportunities for new stochastic questionnaire data modeling and the construction of "instant" latent variables, the same problems caused by lack of attention to measurement and meaning remain. I find I am in agreement with many of the author's views and arguments, but I also find myself wondering whether modern psychometrics and individual differences research methods is now so dominated by psychological statisticians that any thought of substantive scientific innovation in this area that deals more properly with measurement and meaning is long gone. I think the answers to the question "What if there was no psychometrics?" would be most illuminating.

In "Conceptual Complexity and Construct Validity," McGrath (2005/this issue) provides some much needed exploration of processes by which psychologists define and assess constructs and the variables that are said to be constitutive of them. There are, for me, some very significant observations and statements made by McGrath throughout the article that I address in more detail.

QUANTITATIVE AND NONQUANTITATIVE SCIENCE

In the very first sentence of the article, McGrath (2005/this issue) states "*One of the basic requirements of science is accurate measurement* [italics added]" (p. 112). I think this statement (and some others in the article that equate science with quantitative measurement) needs some further elaboration and clarification.

The idea that for anything to be considered scientific, it must somehow involve quantitative measurement, has evolved from Pythagoras (approximately during the 6th century BC). Pythagoras's philosophy stated that nature and reality itself was revealed through mathematics and numerical principles. These numerical principles were proposed as explaining psychological as well as physical phenomena. Given that mathematics might provide the principles by which all phenomena might be understood and given it can be considered the science of structure (Parsons, 1990; Resnick, 1997), then it is reasonable to assume that mathematics could indeed be the means by which nature and reality might be understood. This was the driving philosophy behind the scientific revolution in the 17th century. With the success of quantitative physics in the 19th century came an almost absolute certainty that what could not be measured was of no substantive scientific import. The Kelvin dictum was born during that century (Thomson, 1891):

> I often say that when you can measure what you are speaking about and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge but you have scarcely in your thoughts advanced to the stage of science, whatever the matter may be. (pp. 80–81)

It is also illuminating to read a quotation from Kelley in 1929 summing up the position that intelligence is a measurable variable:

> Our mental tests measure something, we may or may not care what, but it is something which it is to our advantage to measure, for it augments our knowledge of what people can be counted upon to do in the future. The measuring device as a measure of something that it is desirable to measure comes first, and what it is a measure of comes second. (p. 86)

The problem with the original and neo-Pythagorean views is that they assume that all structures, entities, and phenomena can be described by the mathematics of quantity using quantitatively structured variables. That much of the natural sciences could be described in this manner was taken as the signal that psychological constructs could be similarly measured, albeit with some initial difficulty. The original philosophy of Pythagoras had been distorted in the 17th through 19th centuries into what Michell (1999) called the "quantitative imperative." If a discipline could not demonstrate measurement of its constructs and variables, then it could not be considered a science. However, science is a method or process for the investigation of phenomena. It does not require that the variables within its domain of enquiry be quantitatively structured. Quantitative science, in contrast, does demand such properties of its variables. Therein lies the simple yet fundamental distinction between a quantitative science and a nonquantitative science. As Michell (2001) pointed out, there is no preordained necessity for variables within psychology to possess a quantitative structure. Psychology may remain a science yet deal with both quantitative and qualitative (nonquantitative) variables. Quantity is not synonymous with mathematics. If mathematics is considered as the science of abstract structure, then it is obvious that not all structures studied using mathematics are quantitative. For example, the structure of communication and social networks, graphs, language grammars, therapeutic interactions, automata networks, and so forth are essentially nonquantitative. The study of them may remain scientific in that the method of investigation and critical reasoning is applied in accordance with scientific principles, but the variables may be composed of a mixture of the quantitative and nonquantitative or solely nonquantitative. Further discussion of these issues and the substantive consequences of better understanding the logic of measurement rather than psychometric test theory are provided in Barrett (2003, 2005).

CONCEPT COMPLEXITY AND BRUNSWIK (1952) SYMMETRY

Within the section titled "Problems With Complex Constructs" of McGrath's (2005/this issue) article, he notes that "*Studies that have compared scales reflecting constructs at different levels of complexity consistently find that prediction is enhanced by using a larger number of more specific personality variables rather than a smaller number of more global ones* [italics added]" (p. 114). I think Wittmann's (1988; Wittmann & Süb, 1997) concept of "Brunswick Symmetry" based on the "lens" model introduced by Brunswik (1952) is important to introduce at this point, as it greatly clarifies thinking about the level of conceptual complexity of a predictor in relation to the level of complexity of a variable to be predicted. The Brunswik lens model assumes that there are three basic components in any decision-making event: a decision maker, information, and a decision criterion. The model describes the relations among the attributes of a stimulus object (the decision criterion), cues in the environment (information), and the judgement made by a perceiver (the decision maker). The extent to which a decision might be adjudged as accurate is dependent on how well the judgmental use of information equates with the actual value of the environmental cues/information. In other words, the more proximal the decision maker's capacity for utilizing information is to the "objective" relationship of the information to the criterion, the better will be the decision. Wittmann and Süb extended this conceptualization to include symmetry relations between hierarchical predictor and criterion-variable sets such as those that might exist between variables that constitute a hierarchically organized model of personality or intelligence (the predictors) and those that constitute hierarchically organized elements of job performance. Figure 1, taken from Wittmann (2003), displays the schematic features of the Brunswik symmetry model.

What this highlights is that the strongest possible relations between any two variables will be found where perfect level symmetry exists between them. So, for example, correlating the high-level predictor variable PR_g with criterion variable CR_g would be expected to yield a far higher value than if corre-

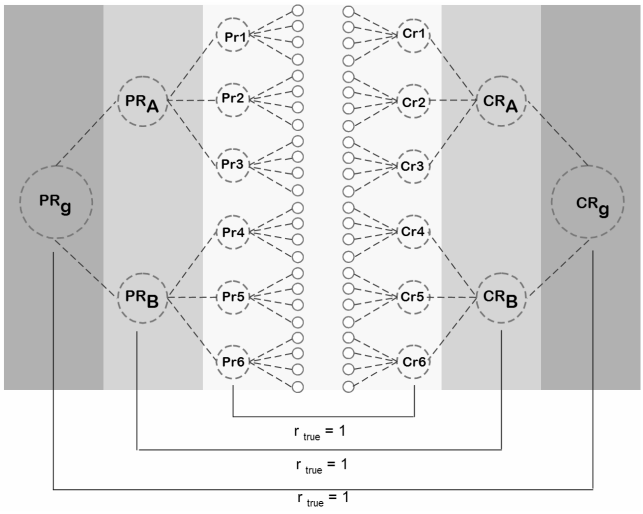

FIGURE 1 The Brunswik symmetrical latent structure of nature (Wittmann, 2003). Reprinted by permission.

lating PR_g with *CR6.* In terms of McGrath's (2005/this issue) exposition of conceptually complex variables, one might view the Wittmann (2003) framework as one potential way of explicating variable complexity and the putative relations between constituent components of a complex variable and other variables. In short, the reason why some predictions fail when using complex (multiconstituent) variables is not just that the complexity of a global variable may be disguising constituents that share little in common with the overall global measure but that the criterion itself has been so badly construed such that the symmetry between predictor and criterion is not what it should be. However, Brunswik (1952) symmetry is not a panacea for a construct whose components are poorly constructed and for one in which no coherent theory exists about the nature and constituent properties of that construct. This point is brought home clearly by McGrath in his discussion of Revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992) facet scales, especially the feelings subscale to which he refers explicitly.

CROSS-MODAL CONSTRUCT CONSISTENCY

With respect to McGrath's (2005/this issue) comments on cross-modal consistency of constructs, especially with regard to "*the common assumption that a self-report measure can sufficiently represent a complex construct such as anxiety* [italics added]" (p. 115), Weinberger, Schwartz, and Davidson's (1979) work on the cognitive processes involved in anxiety and M. W. Eysenck's (1997a, 1997b) subsequent extension and theory of anxiety demonstrated the fallacy that a self-report measure of anxiety is an accurate magnitude of a single unidimensional variable. The expectations and theory of H. J. Eysenck (1967) that magnitudes of self-report anxiety were caused (or depended on) functioning of the limbic system or Gray's (1982) theory that implicated septohippocampal system functioning have been shown to be largely unfounded in a review of the evidence by Fahrenberg (1987, 1992). The problem for researchers who take a low score on a self-report anxiety measure as indicative of "low trait anxiety" is that this group are in fact heterogenous given the meaning of an anxiety latent variable is defined by the correspondence of self-report and physiological response indexes. Weinberger et al. (1979), in an experiment designed to show this disparity of construct criteria, obtained measures of trait anxiety and social desirability from a group of participants in an experiment designed to test their theory that high scores on social desirability confound the self-report measure of anxiety such that low scorers on an anxiety scale and a social desirability scale might genuinely possess low levels of trait anxiety, or, if they obtain a high score on social desirability but a low score on anxiety, they are in fact "repressors." That is, their physiological response will be more similar to a high-anxiety individual, yet their cognitive repression of their own elevated trait anxiety level causes them to be assessed as "low-anxiety" individuals on any questionnaire measure. To test this theory-based hypothesis, the participants were exposed to a moderately stressful situation. The repressors responded physiologically much more than the truly low-anxious individuals. Of great significance is that the repressors responded physiologically at magnitudes beyond individuals who possessed high self-report trait levels of anxiety. This work, as M. W. Eysenck (1997a) indicated, has been replicated by independent investigators. Not only does this finding have enormous implications for the acquisition of evidence for the Gray and H. J. Eysenck models of anxiety (and explains Fahrenberg's [1992] disappointing conclusions in his review of the evidence), but with respect to McGrath's point concerning the complexity of psychological constructs, it reinforces his arguments considerably. It is also interesting to note that construction of a latent variable of anxiety using item response theory (IRT; Ferrando, 2003) would be as ineffective as classical test theory given the same self-report data. I say more about this following.

CONSTRUCT VALIDITY AND MEASUREMENT

Under the heading of "Multi-Item Scales as Representations," McGrath (2005/this issue) states "*The bias toward criterion-related validity as an evidentiary basis for construct validity leads test users to assume the larger correlations associated with multi-item scales means that they are better scales overall* [italics added]" (p. 117). McGrath goes on to note that what makes a good predictor might have no association with what is necessarily a good representation of the construct, and therein lies the heart of the issue—How can we detect if our conceptualization of a construct, embodied in a measurement process, is valid? This question and its component parts are similar to the question posed by Borsboom, Mellenbergh, and van Heerden (2004) in an article devoted to the concept of validity. Borsboom et al. proposed a simple definition of validity:

> If something does not exist then one cannot measure it. If it exists but does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether. Thus a test is valid for measuring an attribute if and only if a) the attribute exists, and b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. (p. 1061)

Of course, hidden in this disarmingly simple definition of validity, in fact, "construct validity," is the phrase "causally produce variations." What this requires is what McGrath (2005/this issue) also asks for some deep consideration of what the properties of an attribute might be so that an investigator might recognize its causal effects when using a test that purports to measure magnitudes of it. A critical point though,

made by Borsboom et al. (2004), is that developing tests based on their optimized correlations with external criteria is unlikely to produce tests that possess valid measurements. Correlations between tests and criteria may be indicative of a causal relation but by themselves cannot constitute a claim of validity.

Perhaps the clearest exposition of the problem associated with the last 50 years of validity theory is that provided by Maraun (1998) who argued that the Cronbach and Meehl (1955) doctrine of construct validity is deeply flawed. Cronbach and Meehl stated

> Scientifically speaking, to "make clear what something is" means to set forth the laws in which it occurs. (p. 290)

Maraun's response to the Cronbach and Meehl statement was

> This is mistaken. One may know more or less about *it,* build a correct or incorrect case about *it,* articulate to a greater or lesser extent the laws into which *it* enters, discover much, or very little about *it.* However, these activities all presuppose rules for the application of the concept that denotes *it* (e.g. intelligence, dominance). Furthermore, one must be prepared to cite these standards as justification for the claim that these empirical facts are about *it.* … [T]he problem is that in construct validation theory, *knowing* [italics added] about something is confused with an understanding of the *meaning* [italics added] of the concept that denotes that something. (p. 448)

So, as with the many linear covariance and item response models that invoke concepts of personality and intelligence as causal latent variables associated with certain phenomena, the knowledge is bound up in the numeric operations applied rather than in the meaning of what actually constitutes an intelligence or personality variable. This is a subtle but telling mistake that becomes apparent when an investigator is asked to explain what it is that the observed test scores are said to be a measurement of and how such a cause might come to possess equal interval and additive unit magnitude relations.

As you read these three contributions, McGrath (2005/this issue), Borsboom et al. (2004), and Maraun (1998), a common theme can be seen to emerge. That is, the generation of a construct or latent attribute requires some initial thinking about what it might be, what might be its constituents, and how any measure proposed for it (or them) might reflect magnitudes of it (or them). This first requires substantive thought about the proposed meaning of such a variable or attribute and then some serious attention paid to how it might be measured validly given that initial set of meaning propositions. The links between the meaning and the measurement must be in place before any test is generated. That much is clear from all three contributions. Furthermore, note another statement from Maraun:

> The relative lack of success of measurement in the social sciences as compared to the physical sciences is attributable to their sharply different conceptual foundations. In particular, the physical sciences rest on a bedrock of technical concepts, whilst psychology rests on a web of common-or-garden psychological concepts. These concepts have notoriously complicated grammars [of meaning]. (p. 436)

This is exactly what McGrath (2005/this issue) is referring to in the many parts of his article in which he discusses the complexity of concepts proposed within the domain of personality and individual differences. Maraun's (1998) position was that whatever measurement is to be created, if at all possible, will need to be created within a normative frame of meaning, with constructs defined in such a specific way as to permit this. That is, it is impossible to create measures of intelligence or depression unless these constructs/phenomena have a normative meaning such that all investigators can work within this common semantic framework. Without this normative agreement, as is largely the case today, tests and measures are produced with no common units and ambiguous normative meaning. A simple example of this is the recent article by Salgado (2003) on the differential predictive validities of tests that purport to measure constructs from the Five-factor model versus those such as the NEO–PI–R, which are said to be constitutive of it. For example, from a collection of tests that all purport to measure conscientiousness, they differ markedly in their prediction of very similar criteria (as well as in some cases possessing correlations far less than 1.0 between their scores). Operational validity is then assigned to the NEO/Five-factor model tests on the basis that the scores on these correlate higher with an outcome than do the others. There is of course no consideration at all in Salgado's article of construct validity. Criterion maximization is associated implicitly with test validity.

CONSTRUCT REPRESENTATION IN A VACUUM OF CAUSAL THEORY

McGrath (2005/this issue) states an important purpose of measurement in the second paragraph of the section of his article titled "Prediction and Representation": "*A measure can also be used as a representation of a construct. This occurs when the measurement is primarily intended to reflect an individual's location on the construct that ostensibly underlies the measure* [italics added]" (p. 113). Latent variable theory and its embodiment especially with IRT would seem to be the ideal approach to generating such constructs from questionnaire items sets. The whole basis of IRT is to establish a functional relationship between item responses, individuals, and a latent variable such that both persons and items may be located on that latent variable. Its measurement properties are assumed to be continuous, linear, and thus possessing the properties of an additive unit concatenation. Unfortunately, even when using the Rasch model (Rasch, 1960), which is perhaps the strongest data model for a linear latent trait vari-

able, the constructed latent variable may bear no relation to any useful, substantive, or even meaningful psychological variable. This was shown quite clearly by Wood (1978) when he fit (very well in fact) random coin tosses with the Rasch model, producing a latent variable of "coin tossing ability." Furthermore, a recent article by Michell (2004) demonstrated quite clearly that the construction of an IRT latent variable conveys nothing about any relationship between it and a hypothesised, empirical, quantitatively structured latent variable. This again is the notion of the causal operations being required to be made explicit for the proposed magnitude changes in any variable. Merely constructing data models to explain item responses might have some pragmatic value, but this is likely to be modest given Fan's (1998) empirical work showing that classical test theory and IRT parameters barely differ in large, representative-sample data sets, contrary to the "received wisdom" of the superiority of IRT over classical test theory. Furthermore, Michell (2004) noted the paradox that as the measures of an attribute are made more accurate (in that less and less measurement error is associated with them), so the IRT models will increasingly fail to fit the data. If one was able to make perfect, error-free measures of levels of an attribute within individuals, no IRT model would fit these data. It is clear that the entire quantitative IRT edifice rests on the requirement that the observations submitted to the model possess a certain amount of error. This is why the entirely random coin-toss data from Wood fits so well. It is in fact the error than permits an IRT model to be fit. It is unclear that any other science strives so hard to maintain error in its observations to subsequently attempt to explain phenomena. There are also many other reasons showing that latent variable theory as currently espoused by many psychologists and social scientists within structural equation modeling is equally deficient in terms of construct validation (see Borsboom, Mellenburgh, & van Heerdon, 2003).

Even if one studiously ignores the content of the preceding paragraph, one is left with the other problem of construct validity, that of choosing between alternate representations of a complex construct that are defined solely in terms of a methodology applied to the same data. When there is no theory that meaningfully (causally) links the proposed functional relations between attributes and the measures of them, then one is faced with representational ambiguity and thus no means of exploring construct validity except by subjective preference. The edited book by Plutchik and Conte (1997) on circumplex models of personality and emotions is full of excellent chapters demonstrating that linear trait conceptualizations of personality structure are merely artificial mathematical divisions of a construct space that is circular and what might be called "near fractal" in form. That is, personality attributes might indeed be best represented as coordinates on the perimeter, or radially extending from, the center of a circle. As one tries to isolate an attribute as a unique entity, one can in fact subdivide its meaning into even smaller more detailed components, and so on, very much like the concept of fractal geometric representations. Maraun (1997) demonstrated the completely different representations that are possible for the Five-factor model when using nonmetric multidimensional scaling (which results in the "Big Two" instead of the Big Five [Goldberg, 1990]). How do we choose? Which is the most valid representation of the personality construct space? Given IRT models now exist for the latent variables of the Big Five, are these the best representations of exactly the same data? The only reason psychologists can continue to debate these issues is because the fixation is on modeling data and not on measuring constructs that possess a "meaning laden" theory as to their cause and that use measures that are constructed on the basis of that theory. The following is from Borsboom et al. (2004):

> This may be one of the few instances in which psychology may actually benefit from looking at the natural sciences. In the more exact quarters, nobody starts constructing measurement instruments without the faintest idea of the processes that led to the measurement outcomes. And it is interesting to note that the problem of validity appears never to have played the major and general role it has played in psychology. (p. 1067)

Interestingly this same theme arises in another guise when onsiderations of null hypothesis significance testing are critically examined. As Gigerenzer (2004) and Gigerenzer, Krauss, and Vitouch (2004) noted, no eminent psychologist of the stature of Piaget, Pavlov, Skinner, or Bartlett relied on significance testing to develop their fundamental theories. Instead, attention to theory, meaning, and empirically based observational and experimental work formed their modus operandi. Look at what Blinkhorn (1997) noted in a review of the last 50 years of test theory:

> Contemporary test theory, with its emphasis on statistical rather than psychological models, has become inaccessible to the majority of test users, and predominantly reflects educational rather than psychological concerns. Real progress may depend on the emergence of a new and radical reconceptualization. (p. 175)

Again, the same theme of the scientific sterility of current psychometric statistical models is present in Blinkhorn's review article. It is not that the models are deficient in any way for the kinds of data that may be presented to them. Rather, it is the data themselves that are deficient in meaning. So, application of the models to such data produces variables and variable relations without ever addressing the question of why those particular variables can be found in the first place and how magnitude variations are caused. I do not deny these models might have good pragmatic utility, but this modeling is invariably constrained by an adherence to null hypothesis significance testing and linear quantitative assumptions. As Breiman (2001) and Gigerenzer (2004) argued, these are but a small part of a statistician's toolbox nowadays and an in-

creasingly minor part as the new algorithmic methods present new knowledge and ways of thinking about, and investigating, phenomena.

SOME PRESCRIPTIVE SPECULATIONS

So, what is one to do? McGrath (2005/this issue) suggests some sensible ways forward that I think many readers might applaud. However, as I hope is apparent from the arguments and many references previously, I think personality and individual differences psychology is going to have to rethink the entire basis on which concepts and constructs have been generated in the past. Theory and empirical evidence from evolutionary psychology and computational and biophysical neuroscience have required a new appreciation of the evolution, development, functioning, and adaptability of the human brain. This knowledge coupled with that from computational evolved systems and population attribute modeling has also allowed researchers to develop the kinds of experimental manipulations and observations that were once thought impossible. For example, instead of academically debating the role and meaning of a trait or disposition such as altruism, I could now use the NetLogo computational modeling software (Wilensky, 1999) to empirically evolve hundreds of generations of populations who live and die, passing on hypothesized genetic predispositions for certain behaviors and who differ in initial levels of altruism and selfishness while also varying external environmental parameters. I do this so that I might test my understanding of the meaning of the trait, its causal basis, its action and dynamics in a population, and the plausibility of that understanding given the operationalization of the variables I consider theoretically related to it. To work with such models, the meaning of altruism has to be specified so that the evolutionary model might be constructed. This is but one way of approaching the construct validity problem that differs markedly from the usual "here are 20 items, let me call them a measure of *X,* and now let me construct a nomological net for them." All one generally finds out in these scenarios is that *X* correlates with almost everything (Meehl's, 1990, "crud" factor), and by reference to the criterion correlations, one infers that *X* is indeed the construct one so defined. At no point has one ever proposed how *X* exists such that one might proceed to observe magnitudes of it, let alone proposing why those magnitudes should be quantitative or nonquantitative.

The challenge that faces many personality and individual differences psychologists is not in understanding or even accepting the reality of McGrath's (2005/this issue) arguments and reasoning—but in figuring out what to do next. I do not underestimate the magnitude of that task. Entire professional careers have been and are continuing to be built on what may well be an illusion that a nomological net equates to construct validity. I am reminded of the devastatingly simple question asked recently by Harlow, Mulaik, and Steiger (1997) by their book title: *What if There Were No Significance Tests?* I think McGrath's arguments suggest psychologists need to consider a new title in exactly the same vein: "What if there were no psychometrics?" With a return to a concern for empirical phenomenal observations; the construction of latent variables based on some kind of meaningful, causal theory for their instantiation; and problems of measurement treated as such rather than problems in stochastic data models or inferential statistics, I think psychologists might just take the next big step in the evolution of psychological science whether or not it is a quantitative science.

REFERENCES

Barrett, P. T. (2003). Beyond psychometrics: Measurement, nonquantitative structure, and applied numerics. *Journal of Managerial Psychology, 3,* 421–439.

Barrett, P. T. (2005). Person-target profiling. In A. Beauducel, B. Biehl, M. Bosniak, W. Conrad, G. Schönberger, & D. Wagener (Eds.), *Multivariate research strategies: A festschrift for Werner Wittmann* (pp. 63–115). Aachen, Germany: Shaker Verlag GmbH.

Blinkhorn, S. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 175–186.

Borsboom, D., Mellenburgh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16,* 199–231.

Brunswik, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO–R) and NEO Five-Factor Inventory (NEO–FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

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

Eysenck, H. J. (1967). The biological basis of personality. Springfield, IL: Thomas.

Eysenck, M. W. (1997a). *Anxiety and cognition: A unified theory.* Hove, England: Psychology Press.

Eysenck, M. W. (1997b). Anxiety and cognitive processes. In C. Cooper & V. Varma (Eds.), *Processes in individual differences: A festschrift in honour of Paul Kline* (pp. 59–72). London: Routledge.

Fahrenberg, J. (1987). Concepts of activation and arousal in the theory of emotionality (neuroticism): A multivariate concept. In J. Strelau & H. J. Eysenck (Eds.), *Personality and dimensions of arousal* (pp. 99–120). New York: Plenum.

Fahrenberg, J. (1992). Psychophysiology of neuroticism and emotionality. In A. Gale & M. W. Eysenck (Eds.), *Handbook of individual differences: Biological perspectives* (pp. 179–226). Chichester, England: Wiley.

Fan, X. (1998) Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological measurement, 58,* 357–381.

Ferrando, P. J. (2003). The accuracy of the E, N, and P trait estimates: An empirical study using the EPQ–R. *Personality and Individual Diffferences, 34,* 665–680.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33,* 586–606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask.

In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Gray, J. (1982). *The neuropsychology of anxiety.* Oxford, England: Clarendon.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kelley, T. L. (1929). *Scientific method.* Chicago: Ohio State University Press.

Maraun, M. D. (1997). Appearance and reality: Is the Big Five the structure of trait descriptors?. *Personality and Individual Differences, 22,* 629–647.

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8,* 435–461.

McGrath, R. E. (2005/this issue). Conceptual complexity and construct validity. *Journal of Personality Assessment, 85,* 112–124.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

Michell, J. (1999). *Measurement in psychology: Critical history of methodological concept.* Cambridge, England: Cambridge University Press.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36,* 211–217.

Michell, J. (2004). Item response models, pathological science, and the shape of error. *Theory and Psychology, 14,* 121–129.

Parsons, C. (1990). The structuralist view of mathematical objects. *Synthese, 84,* 303–346.

Plutchik, R., & Conte, H. R.(Eds.). (1997). *Circumplex models of personality and emotions.* Washington, DC: American Psychological Association.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Resnick, M. D. (1997). *Mathematics as a science of patterns.* Oxford, England: Clarendon.

Salgado, J. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76,* 323–346.

Thomson, W. (1891). *Popular lectures and addresses* (Vol. 1). London: MacMilllan.

Weinberger, D. A., Schwartz, G. E., & Davidson, J. R. (1979) Low-anxious, high-anxious, and repressive coping styles: Psychometric patterns and behavioural and physiological responses to stress. *Journal of Abnormal Psychology, 88,* 369–380.

Wilensky, U. (1999). NetLogo [Computer software]. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University, Retrieved from http://ccl.northwestern.edu/netlogo

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: Plenum.

Wittman, W. W. (2003). *How to fool yourself with experiments in testing theories in program evaluation and psychological research.* Retrieved August 8, 2005, from http://www.psychologie.uni-mannheim.de/psycho2/psycho2.en.php3?page=publi/papers.en.htm&cat=publi

Wittmann, W. W., & Süb. (1997, July). *Challenging G-mania in intelligence research: Answers not given, due to questions not asked.* Paper presented at the ISSID Conference, Aarhus, Denmark. Retrieved August 9, 2005, from www.psychologie.unimannheim.de/psycho2/psycho2.en.php3?page=publi/papers.en.htm&cat=publi

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

Paul Barrett
Department of Management and Employment Relations
Commerce Building C
18 Symonds Street
Auckland 1 New Zealand
Email: paul.barrett@auckland.ac.nz

Received February 19, 2005