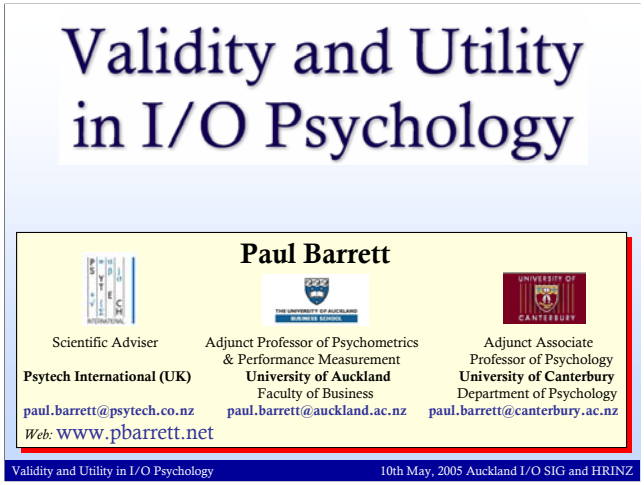
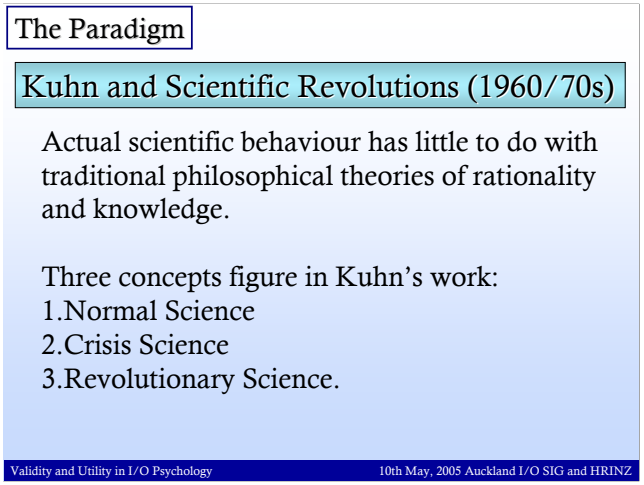


This presentation seeks to offer a sample of several fairly recent major publications and abstracts from a variety of authors which seem to indicate that the current paradigm of psychology and psychometrics is beginning to "dissolve". It is not for me to exhort the audience to be convinced or accept my own judgement as veridical on these sets of evidence, argument, logic, and informed opinion. However, they do seem to ask major questions of academic psychologists as well as applied psychologists as to whether some or all of the below is valid or even partially valid. Individuals like myself have responded to many of these issues with a critical acceptance and complete change of approach to investigating and applying psychological knowledge and methods in research and applied practice. This does not mean we reject what has gone before, but rather we better understand some of its limitations and the limitations of the methods we have been relying upon to get us this far. The problem we face is that some of the now apparent limitations (as with inferential concepts of statistical data models, measurement, and validity) seem sufficiently substantive as to make us wonder as a profession whether we are indeed at the beginning of a paradigmatic rather than evolutionary change. The implications for the future of both academic and applied psychology are huge. But, let's be clear, if change is coming – the history of science teaches us that it will at first be slow, grudging, and divisive.



Normal Science

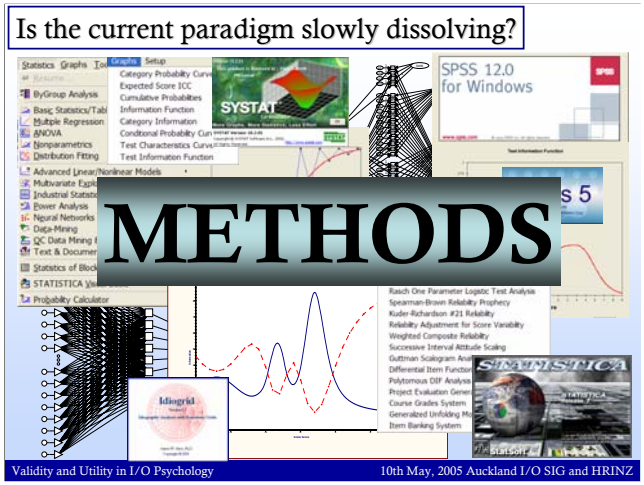
That which proceeds on a daily basis – within a socially agreed and normative paradigm. (a paradigm is a package of claims about the world, methods for gathering and analyzing data, and habits of scientific thought and action.) It is well organized, its knowledge is acquired incrementally.

Crisis Science

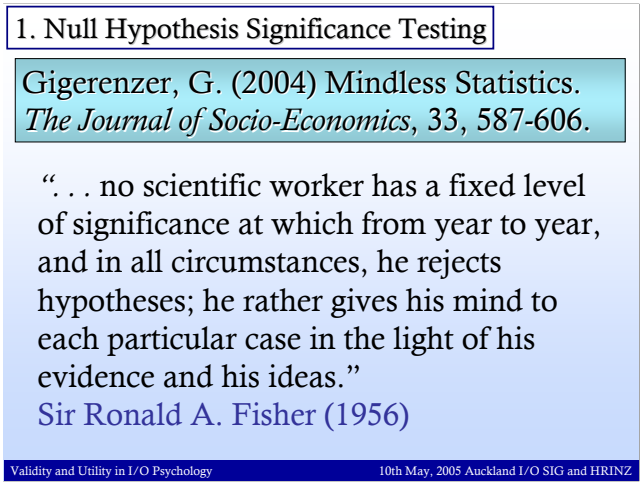
This is what happens when a normal-science paradigm begins to breaks down, **because new observations seem no longer to support the existing paradigm**. However, if there is no new paradigm in existence, the science lurches ahead almost in a "crisis" mode - where the normative paradigm is seen as flawed, but no new one has yet emerged.

Revolutionary Science

Within crisis science, a new paradigm emerges which supplants the old one. It is in effect a "discontinuity" between the "old and the new". However, this process may take years as normal science breaks down into crisis science, with the revolutionary paradigm suddenly emerging/appearing as a nexus of a new normative paradigm.



Here I look at 6 areas in which significant criticisms have been made, but where psychologists and social sciences have mainly chosen to ignore them. Some are devastating to substantive domains of psychological methods. I do not discuss qualitative research methods here as these are confined mainly to guidelines and procedures to aid in observational and qualitative information-gathering exercises. Whilst an essential component for the social sciences, the "knowledge claims" of qualitative research are limited somewhat by the processes required to maintain relatively unconstrained observations and interpretations. Given also the somewhat marginal role of qualitative research methods in much of the social sciences, it is also a moot point whether these could be said to be paradigmatic in the sense that "statistical methods" have now become in these sciences.



Abstract

Statistical rituals largely eliminate statistical thinking in the social sciences. Rituals are indispensable for identification with social groups, but they should be the subject rather than the procedure of science. What I call the "null ritual" consists of three steps:

(1) set up a statistical null hypothesis, but do not specify your own hypothesis nor any alternative hypothesis,

(2) use the 5% significance level for rejecting the null and accepting your hypothesis, and

(3) always perform this procedure.

I report evidence of the resulting collective confusion and fears about sanctions on the part of students and teachers, researchers and editors, as well as textbook writers.

1. Null Hypothesis Significance Testing

Gigerenzer, G. (2004) Mindless Statistics. *The Journal of Socio-Economics*, 33, 587-606.

"Stanley S. Stevens, a founder of modern psychophysics, together with Edwin Boring, known as the "dean" of the history of psychology, blamed Fisher for a "*meaningless ordeal of pedantic computations*" (Stevens, 1960, p. 276). The clinical psychologist Paul Meehl (1978, p. 817) called routine null hypothesis testing "*one of the worst things that ever happened in the history of psychology*.." p. 591

"Rituals seem to be indispensable for the self-definition of social groups and for transitions in life, and there is nothing wrong with them. However, they should be the subject rather than the procedure of social sciences. Elements of social rituals include (i) the repetition of the same action, (ii) a focus on special numbers or colors, (iii) fears about serious sanctions for rule violations, and (iv) wishful thinking and delusions that virtually eliminate critical thinking (Dulaney and Fiske, 1994). **The null ritual has each of these four characteristics: the same procedure is repeated again and again; the magical 5% number; fear of sanctions by editors or advisors, and wishful thinking about the outcome, the *p*-value, which blocks researchers' intelligence.** " p. 603. "

1. Null Hypothesis Significance Testing

Gigerenzer, G. (2004) Mindless Statistics. *The Journal of Socio-Economics*, 33, 587-606.

"If psychologists are so smart, why are they so confused? Why is statistics carried out like compulsive hand washing? My answer is that the ritual requires confusion. To acknowledge that there is a statistical toolbox rather than one hammer would mean its end, as would realizing that the null ritual is practiced neither in the natural sciences, nor in statistics proper. " p. 590

Validity and Utility in I/O Psychology 10th May, 2005 Auckland I/O SIG and HRINZ

"Statistical theory has provided us with a toolbox with effective instruments, which require judgment about when it is right to use them. When textbooks and curricula begin to teach the toolbox, students will automatically learn to make judgments. And they will realize that in many applications, **a skilful and transparent descriptive data analysis is sufficient, and preferable to the application of statistical routines chosen for their complexity and opacity**. **Judgment is part of the art of statistics**. To stop the **ritual**, we also need more guts and nerves. We need some pounds of courage to cease playing along in this embarrassing game. This may cause friction with editors and colleagues, but it will in the end help them to enter the dawn of statistical thinking." p. 604.

2. The Obsession with Data Models

Breiman, L. (2001) Statistical Modeling: the two cultures. *Statistical Science*, 16,3,199-231.

The Data Modelling Culture:

Model validation. Yes–no … using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisticians.

Abstract.

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

2. The Obsession with Data Models

Breiman, L. (2001) Statistical Modeling: the two cultures. *Statistical Science*, 16,3,199-231.

The Algorithmic Modelling Culture:

Model validation. Measured by predictive accuracy.

Estimated culture population. 2% of statisticians, many in other fields.

Validity and Utility in I/O Psychology 10th May, 2005 Auckland I/O SIG and HRINZ

"Misleading conclusions may follow from data models that pass goodness-of-fit tests and residual checks. But published applications to data often show little care in checking model fit using these methods or any other . For instance, many of the current application articles in *JASA* that fit data models have very little discussion of how well their model fits the data. The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model." (p. 203)

"Mosteller and Tukey(1977) were early advocates of cross-validation. They write , "Cross-validation is a natural route to the indication of the quality of any data-derived quantity . We plan to cross-validate carefully wherever we can." Judging by the infrequency of estimates of predictive accuracy in *JASA*, this measure of model fit that seems natural to me (and to Mosteller and Tukey) is not natural to others. More publication of predictive accuracy estimates would establish standards for comparison of models, a practice that is common in machine learning", (p. 204)

3. Structural Equation Modelling

Marsh, H.W., Kit-Tai Hau, Z. Wen (2004) In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling*, 11, 3, 320-341.

Nobody now knows what constitutes acceptable close-fit in Structural Equation Modelling – all we have left is the χ^2 exact-fit test, which rejects nearly every model submitted to it.

Abstract

Goodness-of-fit (GOF) indexes provide "rules of thumb"—recommended cutoff values for assessing fit in structural equation modeling. Hu and Bentler (1999) proposed a more rigorous approach to evaluating decision rules based on GOF indexes and, on this basis, proposed new and more stringent cutoff values for many indexes. This article discusses potential problems underlying the hypothesis-testing rationale of their research, which is more appropriate to testing statistical significance than evaluating GOF. Many of their misspecified models resulted in a fit that should have been deemed acceptable according to even their new, more demanding criteria. Hence, rejection of these acceptable-misspecified models should have constituted a Type 1 error (incorrect rejection of an "acceptable" model), leading to the seemingly paradoxical results whereby the probability of correctly rejecting misspecified models decreased substantially with increasing N. In contrast to the application of cutoff values to evaluate each solution in isolation, all the GOF indexes were more effective at identifying differences in misspecification based on nested models. Whereas Hu and Bentler (1999) offered cautions about the use of GOF indexes, current practice seems to have incorporated their new guidelines without sufficient attention to the limitations noted by Hu and Bentler (1999).

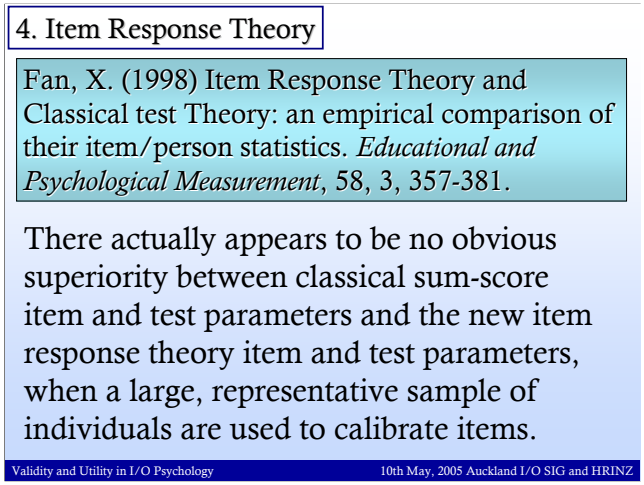
3. Structural Equation Modelling

SEMNET – Prof. Ed Rigdon … Professor of marketing .. Sat 5th March, 2005 … in the SEMNET archives at: http://bama.ua.edu/archives/semnet.html

"The finest thinking about research procedure, about managing Type I and Type II error, about statistical precision and power, all starts from the premise that there's a real price to pay if you're wrong. In many, many applications of SEM, I don't know that there is."

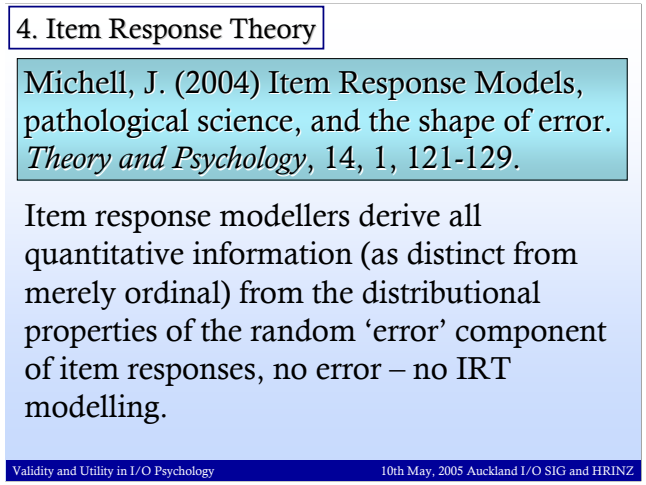
Validity and Utility in I/O Psychology 10th May, 2005 Auckland I/O SIG and HRINZ

"I think that the wobbly feeling is because, in the absence of a real or computable cost of being wrong, the **ritualistic element** of research has taken over. We have so many pieces of a really fine methodology, but we can't answer that final model evaluation question, in my opinion, largely because we can't look back on the decisions we've made, look at the consequences, and link those consequences to our methodological choices. Statistics alone will not resolve the issue, because it's not a statistical issue. **The finest thinking about research procedure, about managing Type I and Type II error, about statistical precision and power, all starts from the premise that there's a real price to pay if you're wrong. In many, many applications of SEM, I don't know that there is."**



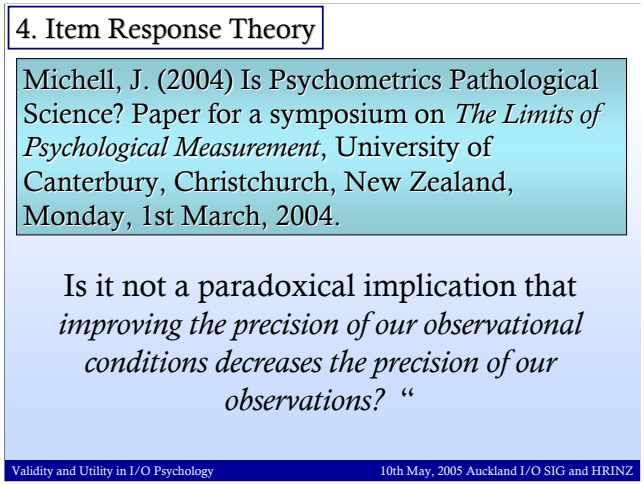
Abstract

Despite theoretical differences between item response theory (IRT) and classical test theory (CTT), there is a lack of empirical knowledge about how, and to what extent, the IRT- and CTT-based item and person statistics behave differently. This study empirically examined the behaviors of the item and person statistics derived from these two measurement frameworks. The study focused on two issues: (a) What are the empirical relationships between IRT- and CTT-based item and person statistics? and (b) To what extent are the item statistics from IRT and those from CTT invariant across different participant samples? A large-scale statewide assessment database was used in the study. The findings indicate that the person and item statistics derived from the two measurement frameworks are quite comparable. The degree of invariance of item statistics across samples, usually considered as the theoretical superiority IRT models, also appeared to be similar for the two measurement frameworks.

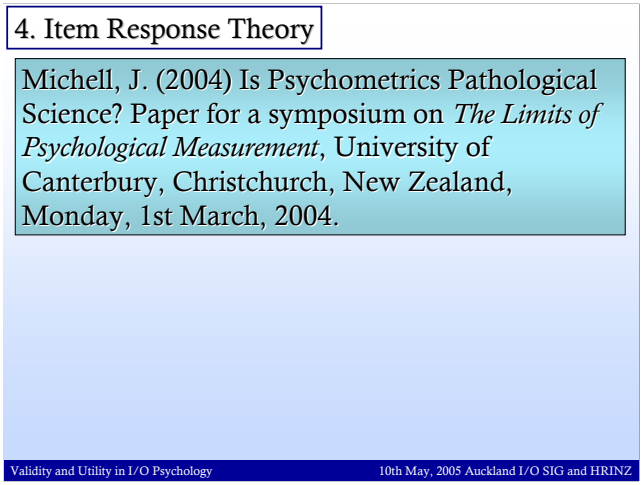


Abstract

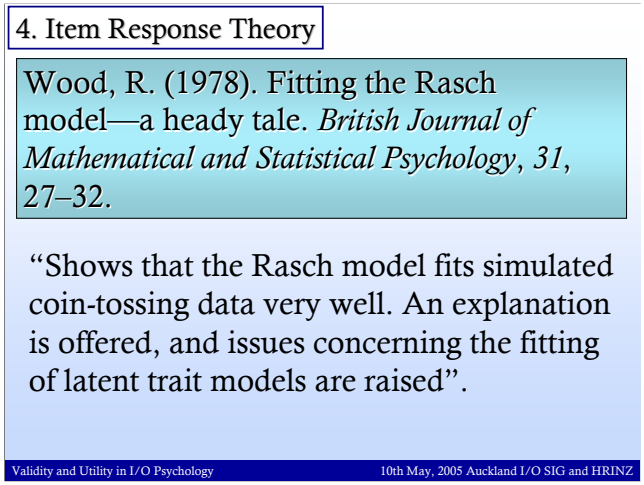
There is nothing in Borsboom and Mellenbergh's (2004) response that refutes my thesis that psychometrics is a pathology of science. They seek to defend item response models from my charge of pathological science without apparently realizing that my charge relates to psychometricians, not to models. They appeal to the Quine–Duhem thesis in an attempt to argue that item response models do not allow the hypothesis that psychological attributes are quantitative to be tested in isolation, but their argument is based upon a misinterpretation of Duhem. In any experiment, what is being tested depends on what the experimenter already takes to be true, and it is possible that a psychometrician could be testing just one of the hypotheses constituting an item response model. Furthermore, using the theory of conjoint measurement, it is possible to isolate predictions that depend upon psychological attributes being quantitative, as opposed to merely ordinal. Despite this, Borsboom and Mellenbergh agree with the first part of my thesis. They do not discuss the second part, but an examination of textbooks on item response models shows that psychometricians disguise their failure to test the hypothesis that psychological attributes are quantitative by simply declining to mention that this hypothesis is presumed in their models. Claims to measure psychological attributes based upon these models depend exclusively upon the weakest part of these models: the hypothesis that the distribution of 'errors' takes a specific form.



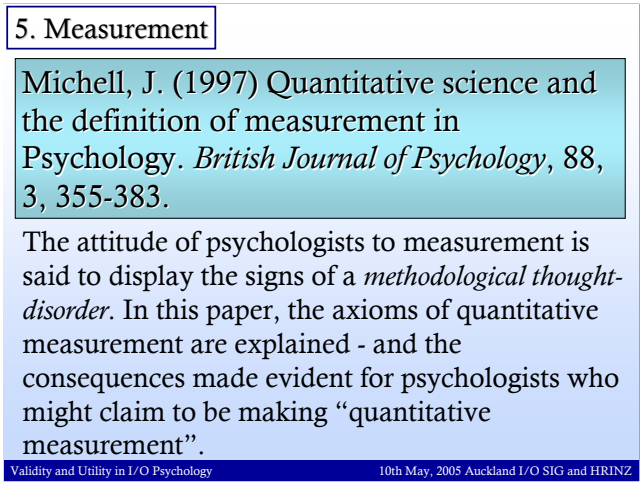
"Guttman's is the simplest of all item response models. In the context of ability tests, it is that a person attempting an item will get it correct if and only if the person's measure on the latent trait is not less than the item's. It is an ordinal model in two senses: first, it only requires that the latent trait have ordinal structure; and, second, it only provides an ordering on people and items on that trait. As Borsboom and Mellenbergh (2004, p.108) note, however, this model is 'very restrictive' in the sense that it fits responses to very few psychological tests. Most modellers interpret this widespread failure as evidence for the existence of *error* in individual differences in test performance. As a result they accommodate *error* in their models and assumptions made about *error*, and these alone, define a model's specifically quantitative structure (Michell, 2004)."



"The place of *error* in parametric models leads to an apparent paradox. The term *error* denotes the effects of factors extraneous to the trait under investigation, which are thought to affect individual differences in responses. Suppose parametric modellers are correct about the relationship between a person's and an item's measures only being discerned through a haze of *error*. Further, suppose that we were able to improve controls in the testing situation and eliminate the effects of extraneous factors, thereby eliminating *error*. This dramatic improvement in the precision of our procedures would lead to no improvement in the precision of our measurements, as such improvements typically would in other sciences. In fact, quite the reverse! Such improvements would mean that we would have only a Guttman scale and, so, would not be able to measure at all, but instead would only be able to put people and items in order, whereas, before, with observations contaminated by *error*, we had quantitative measurement. Is it not a paradoxical implication that *improving the precision of our observational conditions decreases the precision of our observations?* "



Michell continued" … To get this into perspective, think by analogy with procedures in another science, say, astronomy. Suppose we were inspecting some newly discovered star, one that we could only see dimly because of some kind of hazy interference in outer space. Suppose further that through this haze we thought we could detect a system of planets orbiting the star. Then by some lucky circumstance, suppose that the haze disappeared and our view of the star improved and that what we had previously thought was a planetary system could no longer be seen. Would we not feel justified in concluding that what we had thought was a planetary system was really only an artefact of the interference? We would feel this because we are suspicious of effects that depend on error. If things, which we think are there, cannot be detected when the precision of our procedures improves, then we need additional evidence of their existence. Likewise, since the quantitative relationships that we think we can detect via parametric item response models would disappear were *error* eliminated, we require additional evidence of their existence. We need tests that are specifically attuned to the hypothesis that the relevant trait is quantitative. "



Abstract

It is argued that establishing quantitative science involves 2 research tasks: the scientific one of showing that the relevant attribute is quantitative; and the instrumental one of constructing procedures for numerically estimating magnitudes. In proposing quantitative theories and claiming to measure the attributes involved, psychologists are logically committed to both tasks. However, they have adopted their own, special, definition of measurement, one that deflects attention away from the scientific task. It is argued that this is not accidental. From G. T. Fechner (1860) onward, the dominant tradition in quantitative psychology ignored this task. S. S. Stevens's (e.g., 1946, 1951) definition rationalized this neglect. The widespread acceptance of this definition within psychology made this neglect systemic, with the consequence that the implications of contemporary research in measurement theory for undertaking the scientific task are not appreciated. It is argued further that when the ideological support structures of a science sustain serious blind spots like this, then that science is in the grip of some kind of thought disorder.

5. Measurement

Michell, J. (2001) Teaching and misteaching measurement in psychology. *Australian Psychologist*, 36, 3, 211-217.

"This paper argues that the way in which psychometrics is currently, typically taught actually subverts the scientific method."

Abstract

A feature common to all scientific methods is critical inquiry, ie., testing claims. If the teaching of methods in any scientific discipline fails to exemplify this common feature, then the enterprise of science may be subverted. This could happen, for example, if a hypothesis fundamental to some method were not taught critically or, more seriously, if critical scrutiny of it were to be systematically excluded from the curriculum. It is argued here that this has happened in psychometrics with regard to the claim upon which psychological measurement depends, viz., that psychological attributes (eg., intellectual abilities, personality traits, and social attitudes) are quantitative. It is further argued that this has happened because of historical and social pressures within the discipline. Suggestions are made for additions to the measurement curriculum.

5. Measurement

Michell, J. (2005) The Meaning of the Quantitative Imperative: A Response to Niaz. *Theory & Psychology*. 15, 2, 257-263.

Abstract: "In this latter sense, in the historical episode that I have investigated, the quantitative imperative is a political campaign protecting a scientific image of psychology and packaging psychological tests as methods of scientific measurement".

"Within modern psychology, the quantitative imperative has nothing to do with science as the critical investigation of nature's ways of working and everything to do with a political campaign, one serving a number of special, but undeclared, interests. First, this campaign projects an image of psychology as a science, one conforming to a scientistic image derived from the established, quantitative sciences. This was important for the institutionalization of the discipline prior to the First World War and was later important in soliciting financial support in the era of Big Science after the Second World War. Second, it advances the interests of applied psychology (especially psychometrics) by marketing tests as instruments of measurement, packaging them within a ready-made scientific category, one possessing a socially important role. Each of these interests, in turn, derives from obvious social and economic interests, the potency of which cannot be underestimated. However, as a verbal formula, the quantitative imperative does not disclose these interests, making it appear instead that it is the intrinsic character of scientific investigation itself that requires measurement."

6. Meta Analysis

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997) Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *The New England Journal of Medicine*, 337, 8, 536-542.

"The outcomes of 12 large randomized, controlled trials that we studied were not predicted accurately 35 percent of the time by the meta-analyses published previously on the same topics"

Abstract

Methods: We compared the results of large randomized, controlled trials (involving 1000 patients or more) that were published in four journals (the New England Journal of Medicine, the Lancet, the Annals of Internal Medicine, and the Journal of the American Medical Association) with the results of meta-analyses published earlier on the same topics. Regarding the principal and secondary outcomes, we judged whether the findings of the randomized trials agreed with those of the corresponding meta-analyses, and we determined whether the study results were positive (indicating that treatment improved the outcome) or negative (indicating that the outcome with treatment was the same or worse than without it) at the conventional level of statistical significance (P<0.05).

Results: We identified 12 large randomized, controlled trials and 19 meta-analyses addressing the same questions. For a total of 40 primary and secondary outcomes, agreement between the meta-analyses and the large clinical trials was only fair (kappa = 0.35; 95 percent confidence interval, 0.06 to 0.64). The positive predictive value of the meta-analyses was 68 percent, and the negative predictive value 67 percent. However, the difference in point estimates between the randomized trials and the meta-analyses was statistically significant for only 5 of the 40 comparisons (12 percent). Furthermore, in each case of disagreement a statistically significant effect of treatment was found by one method, whereas no statistically significant effect was found by the other.

Conclusions: The outcomes of the 12 large randomized, controlled trials that we studied were not predicted accurately 35 percent of the time by the meta-analyses published previously on the same topics.

7. Disattenuating Correlations

Corrections for Restriction of Range … what's the point?

Disattenuating and interpreting Validity Correlations for restriction of range is only valid for stage **1** in a **2**-stage prediction/classifier process.

Corrections for Restriction of Range … what's the point?

Stage 1: answers the question "what is the correlation between a predictor and criterion given **all possible values** of a predictor and criterion variable might be used as input data"?

Corrections for Restriction of Range and Unreliability … what's the point?

Stage 2: asks the question "what is the correlation between a predictor and criterion given that only a subsample of the entire population of predictor values will ever be **observable** (pre-selection)?"

It is this correlation that practitioners have to work with – not the disattenuated form.

7. Disattenuating Correlations

Corrections for Unreliability … what's the point?

To see what the maximum possible value for a relationship might be if there were no measurement error. This value cannot be used in practice – it yields a hypothetical value that is useful for theory purposes, and for examining the effects of measurement error on a relationship. That's all.

7. Disattenuating Correlations

Predicting Job Performance from GMA - Salgado et al (2003)

Job-Source	Raw Mean Validity	Operational Validity
Driver	0.22	**0.45**
Electrician	0.28	**0.54**
Information Clerk	0.31	**0.61**
Engineer	0.23	**0.63**
Manager	0.25	**0.67**
Police	0.12	**0.24**
Sales	0.34	**0.66**
Skilled Worker	0.28	**0.55**
Typing	0.23	**0.45**

These "operational validity" = disattenuated values are equivalent to those presented in the Hunter and Schmidt meta analysis of 1998..

Schmidt, F.L., & Hunter, J.E. (1998) The Validity and Utility of Selection Methods in Personnel Psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 2, 262-274. – and -

Schmidt, F.L., & Hunter, J. (2004) General mental ability in the world of work: occupational attainment and Job Performance. *Journal of Personality and Social Psychology*, 88, 6, 162-173.

Which just goes to show the problem faced by practitioners when using meta-analytic evidence to try and make sense of job performance within groups of employees within companies. i.e. great for knowing that engineers need a high level of ability as against a mechanic, but not much use for figuring out who is your best vs worst engineer.

Salgado, J.F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J.P. (2003) A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, 88, 6, 1068-1081.

7. Disattenuating Correlations

Job-Source	Raw Mean Validity	Operational Validity
Apprentice	0.26	**0.49**
Chemistry	0.28	**0.72**
Driver	0.26	**0.40**
Electrician	0.35	**0.63**
Information Clerk	0.46	**0.69**
Engineer	0.28	**0.74**
Mechanics	0.21	**0.40**
Police	0.13	**0.25**
Skilled Worker	0.17	**0.27**
Typing	0.31	**0.57**

Predicting **Training Success** from GMA - Salgado et al (2003)

This problem of "if imbeciles as well as geniuses applied for my job" and "I have perfect error-free measurement" reasoning that underpins these corrections is brought home dramatically with regard to Conscientiousness by:

Robertson, I.T., Baron, H., Gibbons, P., MacIver, R., and Nyfield, G. (2000) Conscientiousness and Managerial Performance. *Journal of Occupational and Organizational Psychology*, 73, 2, 171-180.

Abstract

Recent research has provided clear evidence that personality factors are associated with job performance. The construct of conscientiousness has been shown to be a particularly promising predictor of overall job performance. Some authors have proposed that conscientiousness might be the 'g' of personality and predict performance in most occupational areas. The nature of the construct of conscientiousness is reviewed and consideration given to the likely behaviour associated with high conscientiousness. It is hypothesized that, given the requirements of managerial work, the criterion-related validity of conscientiousness may not extend to all managerial jobs. Conscientiousness scores are derived for a sample of managers (*N* = 437), with the aid of personality questionnaire data. In a concurrent validity design these scores are correlated with indicators of current job performance, promotability and specific job performance factors….

7. Disattenuating Correlations

Robertson, I.T., Baron, H., Gibbons, P., MacIver, R., and Nyfield, G. (2000) Conscientiousness and Managerial Performance. *Journal of Occupational and Organizational Psychology*, 73, 2, 171-180.

The overall validity coeffcient obtained by Barrick and Mount (1991) for conscientiousness for managers was **.13** (.22 after correction for range restriction and unreliability).

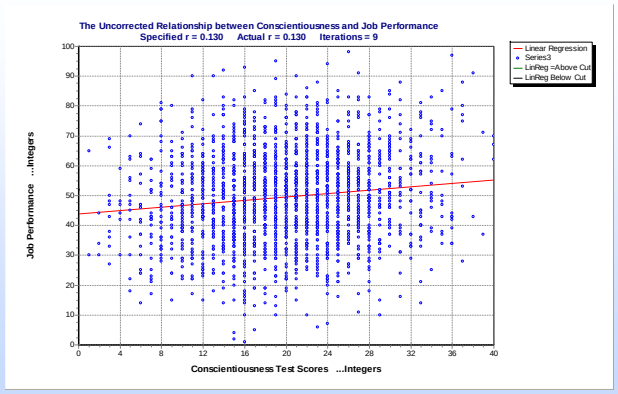
Abstract continued … "The correlation of conscientiousness with current performance is close to zero and the correlation with promotability is -.20. The pattern of relationships between conscientiousness and the job performance factors is used to interpret the finding that conscientiousness is not influential in determining managerial performance. The results suggest that there may be limits to the range of occupational areas in which conscientiousness is closely linked with job performance".

"These findings were consistent across the other occupational groups in their sample, suggesting that conscientiousness is a valid predictor for all occupational areas. Although there has been subsequent controversy about methodological and statistical procedures in their study (Ones, Mount, Barrick, & Hunter, 1994), **Tett *et al*. (1991) found an overall validity for conscientiousness, from confirmatory studies, of .12 (.18 after correction).** Whereas the results from Barrick and Mount (1991) showed generalizable validity for conscientiousness, the results from Tett *et al*. (1991) showed a lower boundary for the confidence interval for conscientiousness of less than zero, leaving the generalizable validity for conscientiousness in doubt and suggesting the existence of possible moderator variables. Using studies conducted in the European Union, Salgado (1997) found values for the average validity of conscientiousness of .06 (uncorrected, .16 after correction). In Salgado's (1997) study, the average validity for managers was lower than other occupational groups (e.g. police showed a corrected coefficient of .39). (p.172)

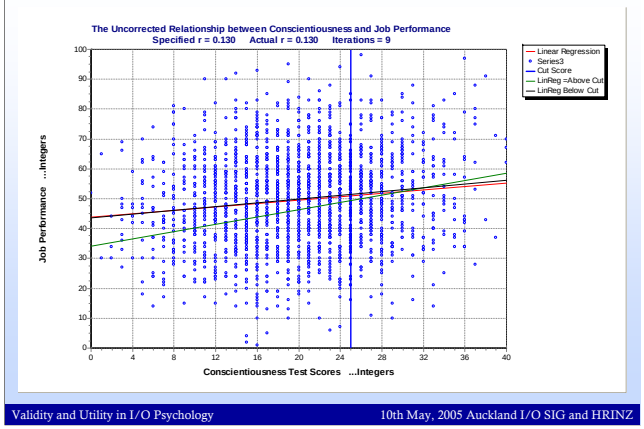
8. Using Correlations

Let's see what the actual data look for these correlations - upon which some I/O psychologists will ask clients to make expensive decisions upon …

8. Using Correlations



8. Using Correlations

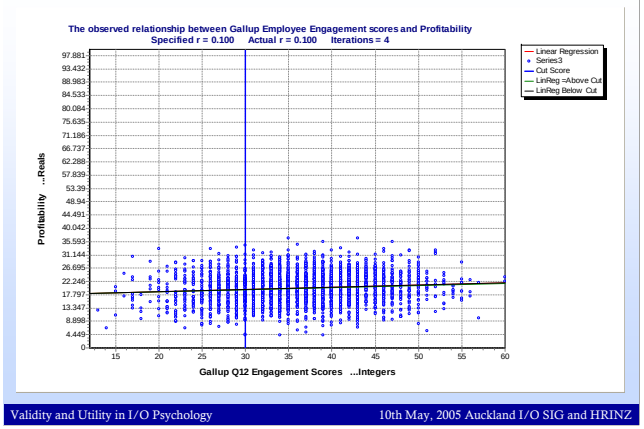


Validity and Utility in I/O Psychology 10th May, 2005 Auckland I/O SIG and HRINZ

Note that the correction for restriction of range in this sample of 5000 cases, with a cut score of 25, yields an over-estimate of the true population value. The green line shows what the correlation would look like as a trend-line in the data – with the red line indicating the population value. Just a reminder that these corrections ASSUME the restricted sample represents EXACTLY the features of the segment of population from which they are drawn. Even with a sample as large as 5000, from which we subselect about 1500 cases, the correction does not hold in this example.

8. Using Correlations

N=200,000 sample

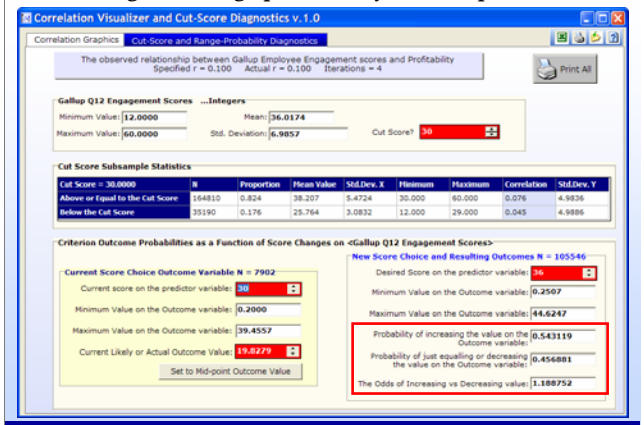


Validity and Utility in I/O Psychology 10th May, 2005 Auckland I/O SIG and HRINZ

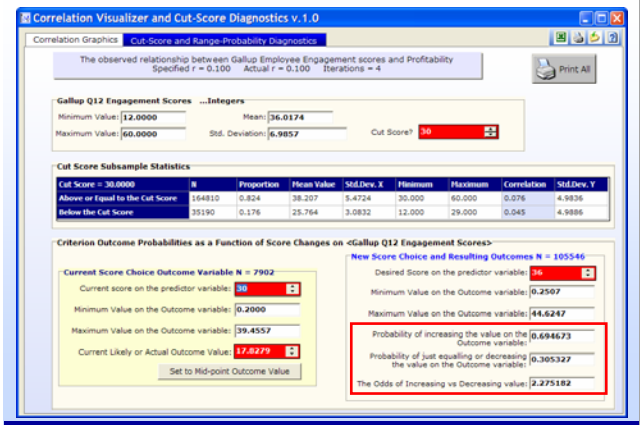
Harter, J.K., Schmidt, F.L., and Hayes, T.L. (2002) Business-Unit-Level Relationship between employee satisfaction, employee engagement, and business outcomes: a meta analysis. *Journal of Applied Psychology*, 87, 2, 268-279.

In a nutshell, the aim of the Gallup organizational survey is to permit item responses to be "benchmarked" against a database of such responses from many organizations. Given a positive relation has been previously empirically determined between the level of response and a positive business outcome, the idea is to determine where your organization stands in relation to others, and then seek to develop intervention strategies that will subsequently lead to an increase in Gallup ratings next year, with the implied benefit that the causes of such ratings changes will indeed lead to more positive business outcomes such as increased profitability. Almost 200,000 employees were used in this meta analysis … and thus 200,000 are used in the simulation above.

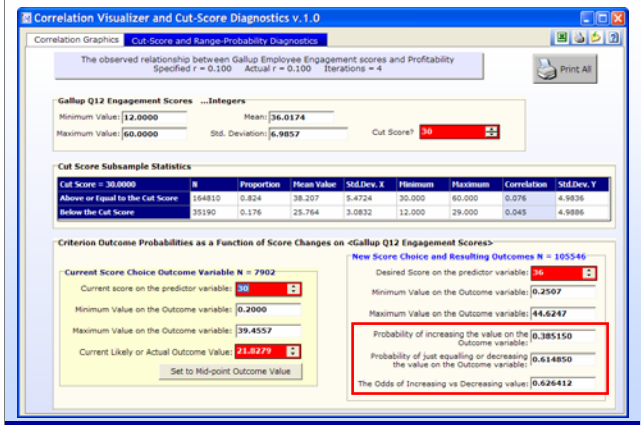
Assuming an average profitability at Gallup score = 30



Assuming a 2% lower-than-average profitability at Gallup score = 30

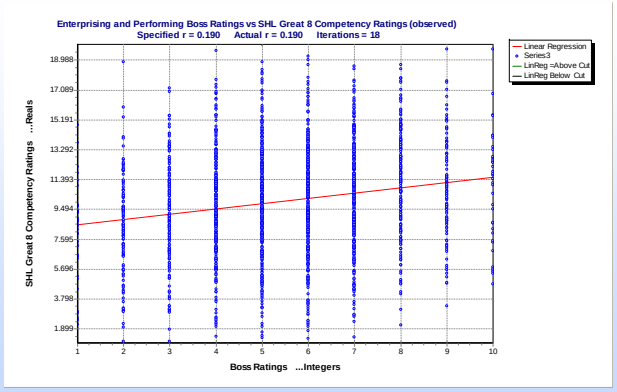


Assuming a 2% higher-than-average profitability at Gallup score = 30



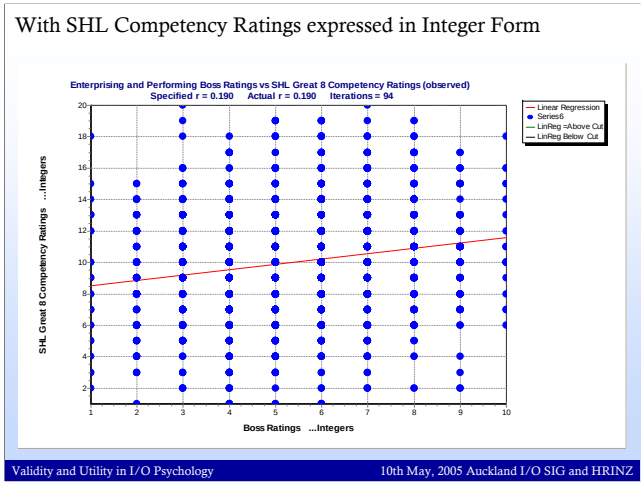
8. Using Correlations

N=4,000 sample

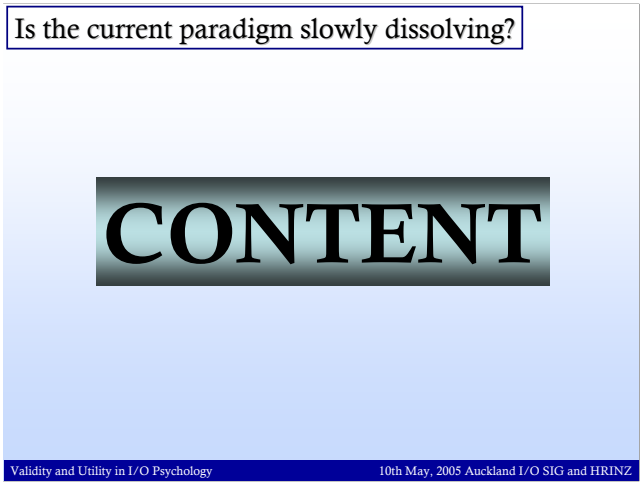


From the SHL Research Paper ..

Bartram, D. (2004) The Great Eight Competencies: A criterion-centric approach to validation. Table 3c.



Abstract
The Great Eight (Kurz & Bartram, 2001) is a model of work-related behaviors that divides the domain of performance at work into eight broad areas. The results of a meta-analysis of 29 validation studies (total N=4861) are presented that uses the Great Eight competency factors as the criterion measurement framework. All the studies had personality data as predictors and line-manager ratings of competencies as criteria. In addition, some of the studies had ability test data and some had measures of overall job performance (OJP). Predictors of the Great Eight competencies based only on personality scales show good correlations with line-manager ratings for all eight of the competencies. On their own ability tests correlate with four of the eight competencies and together, ability and personality data yield corrected sample-weighted **[P.B.note … disattenuated]** correlations ranging from 0.20 to 0.42 for the eight competencies. Moderator analyses show higher average validities for studies which used ipsative (forced-choice) personality instruments rather than normative, concurrent designs rather than predictive, and standardized competency assessment instruments rather than client-based ones. The relationship between the Great Eight predictors and OJP is discussed, and it is noted that the results reflect **the general finding in the meta-analysis literature of a relationship between conscientiousness and job performance**.



Here, I note significant papers and findings that would be practice-changing if only the evidence had not been carefully and studiously ignored for so long. And, don't think this ignorance is a recent phenomena – as you'll see when you read David Lykken's paper, things are much worse than even Joel Michell has stated with specific regard to measurement. What I find curious is that that few (if any) psychologists respond to the evidence or arguments, or are even willing to think about the implications of some of it. Some of the papers presented so far are devastating to the profession as a whole. This has all the hallmarks of a paradigm beginning to fail, being maintained by a majority of psychologists and practitioners as best they can, whilst gradually the new evidence and innovations begin to pop-up more and more as an increasing minority of individuals begin to look beyond current established knowledge-bases and practice.

9. The Big Five

Maraun, M.D. (1997) Appearance and Reality: Is the Big Five the Structure of Trait Descriptors? *Personality and Individual Differences*, 22, 5, 629-647.

As you change the metric of the analysis to better suit personality data (ordinal-level relations and non-metric multidimensional scaling), the dimensionality is reduced to two dimensions and radex structures. So, by changing the method of analysis, you lose the Big Five factors and gain a new Big Two. How do you decide which is a more accurate depiction of human personality?

Abstract

It is argued that contrary to the claims of Big Five investigators, the structure of trait descriptors is still very much an open issue. This is because their methodology, factor/component analysis paired with the dimensional interpretation/simple structure procedure, does not investigate the closed topological manifold that constitutes the "structure" of a set of variables. Instead, radex-related configurations are likely candidates for the structure of trait descriptors. Some preliminary support for this claim is given by an analysis of the NEO Personality Inventory (NEO-PI), a Big Five questionnaire measure, and the Goldberg-40, an adjective measure. The NEO-PI data was the correlation matrix among the 30 NEO-PI facet scales (P. Costa & R. McCrae, 1992). The Goldberg-40 was administered to 215 undergraduates. Results show the NEO-PI and Goldberg-40 have radex structures. A facet theory rationale is provided for these findings

10. Ideational Stagnation in the Test Industry

Sternberg, R.J. and Williams, W. (1998) You proved our point better than we did: A Reply to Our Critics. *American Psychologist*, 53, 5, 576-577).

"No technology of which we are aware- computers, telecommunications, televisions, and so on- has shown the kind of ideational stagnation that has characterized the testing industry. Why? Because in other industries, those who do not innovate do not survive…

10. Ideational Stagnation in the Test Industry

"…In the testing industry, the opposite appears to be the case. Like Rocky I, Rocky II, Rocky III, and so on, the testing industry provides minor cosmetic successive variants of the same product where only the numbers after the names substantially change. These variants survive because psychologists buy the tests and then loyally defend them (see preceding nine commentaries, this issue). The existing tests and use of tests have value, but they are not the best they can be…"

10. Ideational Stagnation in the Test Industry

"…When a commentator says that it will never be possible to improve much on the current admissions policies of Yale and its direct competitors (Darlington, 1998, p. 572, this issue), that is analogous to what some said about the Model T automobile and the UNIVAC computer".

11. Ipsative Tests

Meade, A. (2004) Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, 77, 4, 531-552.

"In sum, the standards required of tests used for employee selection are quite strict with regards to validity and reliability of the selection instruments. **As such, the limitations inherent with ipsative measures pose too great a threat to the validity of the selection tools to make it a useful instrument for selection on a trait-by-trait basis**". p. 549.

Abstract

Data are described as ipsative if a given set of responses always sum to the same total. However, there are many properties of data collection that can give rise to different types of ipsative data. In this study, the most common type of ipsative data used in employee selection (forced-choice ipsative data; FCID) is discussed as a special case of other types of ipsative data. Although all ipsative data contains constraints on covariance matrices (covariance- level interdependence), FCID contains additional item-level interdependencies as well. The psychological processes that give rise to FCID and the resultant psychometric properties are discussed. In addition, data from which both normative and ipsative responses were provided by job applicants illustrate very different patterns of correlations as well as very different selection decisions between normative, FCID and ipsatized measures.

11. Team-Working not so optimal?

Allen, N.J., Hecht, T.D. (2004) The 'romance of teams': toward an understanding of its psychological underpinnings and implications. *Journal of Occupational and Organizational Psychology*, 77, 439-461.

The first sentence of the abstract:
"Although advocates of teamwork suggest that teams enhance performance, empirical evidence does not consistently, or robustly, support these claims"

Abstract

Although advocates of teamwork suggest that teams enhance performance, empirical evidence does not consistently, or robustly, support these claims. Still, a belief in the effectiveness of teams—among managers, employees, and the general lay population—seems very strong. What accounts for this 'romance of teams'? In this paper, we offer a psychological answer to this question. We review evidence regarding the actual effectiveness of teams, in order to show that teams are not as effective as many believe them to be, and we argue that the romance of teams stems from the psychological benefits of group-based activity. Specifically, we propose that team members experience both social-emotional, and competence-related, benefits, and we review an eclectic mix of research in support of this claim. We argue that these psychological benefits of teams lead people to assume that teams are 'high performance', thus, causing the romance of teams. Finally, we discuss potential implications of the romance for organizations, researchers, and employees.

12. Walter Mischel on the US NIMH funding cut

Mischel, W. (2005) Alternative Futures for Our Science. *The Observer: American Psychological Society*, March, 18, 3, 15-19.

"For me, the classic partitioning most unnatural and destructive to the building of a **cumulative science** of mind and social behavior is the one that traditionally has split the person apart from the situation, treating each as if it were an independent cause of behavior. How the field deals with this split will significantly influence the future it constructs for itself " (p. 17)

"Historically, the assumption that the person and the situation are independent causes of behavior led to making personality psychology the field devoted to the person apart from the situation. It treated the situation as the error term that needs to be removed or aggregated away. To see the person, you had to remove the effects of the situation, either by making it completely ambiguous, as on a Rorschach inkblot in projective testing, or by getting rid of it on situation-free global measures of what the person is like "on the whole." Consequently, the situation was — and in much current practice still is — deliberately removed or aggregated out to ask about the general effects of persons, regardless of situations. In contrast, much of social psychology became defined as the study of the effects of situations, usually regardless of the kinds of persons in them. So, for each field, the main variables of the other constituted the error variance that needed to be removed. As Leon Festinger said to me 40 years ago while discussing my interest in personality and individual differences, "Your independent variables are my noise." And I told him that his noise was my essence.

The boundaries between personality and social psychology, and between the person and the situation, made little sense to me when Personality and Assessment was published in 1968. They make even less sense now, because treating the person and the psychological situation as independent causes of behavior flies in the face of the reciprocal interactions that our science is finding, of what the cognitive revolution taught us years ago, and of what is again being found in cognitive neuroscience and biology. Therefore with the goal of studying the person and the situation at their natural joints — rather than at their old academic joints — Yuichi Shoda and I and our colleagues for many years have focused on the situation as well as the person jointly, and on their intrinsic interconnections in the head of the perceiver and in the behaviors that are generated in the social world".

13. So many flaws – a veritable litany

Lykken, D.T. (1991) *What's Wrong with Psychology Anyway?* In D.Cicchetti and W.M. Grove. (Eds.). *Thinking Clearly about Psychology. Volume 1: Matters of Public Interest*. University of Minnesota Press. ISBN: 0-8166-19182.

"Surrounded by difficulties and complexities, we have invented comforting "Cargo Cult" rituals, adopted scientistic fads, substituted pedantry for substance, jargon for common sense, statistical analysis for human judgment"

"The present paper is a distillation of the three lectures I have been contributing to Paul's (Meehl) Philosophical Psychology. I offer it here in fond respect for the man who has been my teacher and friend for nearly forty years.

I shall argue the following theses:

(I) Psychology isn't doing very well as a scientific discipline and something seems to be wrong somewhere.

(II) This is due partly to the fact that psychology is simply harder than physics or chemistry, and for a variety of reasons. One interesting reason is that people differ structurally from one another and, to that extent, cannot be understood in terms of the same theory since theories are guesses about structure.

(III) But the problems of psychology are also due in part to a defect in our research tradition; **our students are carefully taught to behave in the same obfuscating, self-deluding, pettifogging ways that (some of) their teachers have employed."**

13. So many flaws – a veritable litany

Lykken, D.T. (1991) *What's Wrong with Psychology Anyway?* In D.Cicchetti and W.M. Grove. (Eds.). *Thinking Clearly about Psychology. Volume 1: Matters of Public Interest*. University of Minnesota Press. ISBN: 0-8166-19182.

"…the problems of psychology are also due in part to a defect in our research tradition; **our students are carefully taught to behave in the same obfuscating, self-deluding, pettifogging ways that (some of) their teachers have employed**".

"I once was outside reviewer on a dissertation from a Canadian university, a rather interesting-sounding study of autonomic responses of psychopaths, neurotic offenders, and normals. I found it impossible to determine how the study came out, however, because there were 75 pages of ANOVA tables, 4th order interactions, some of them "significant" and discussed at wearying length. I suggested that the candidate should be passed since he clearly had been taught to do this by his faculty but that perhaps some of the faculty ought to be defrocked."

13. So many flaws – a veritable litany

Lykken, D.T. (1991) *What's Wrong with Psychology Anyway?* In D.Cicchetti and W.M. Grove. (Eds.). *Thinking Clearly about Psychology. Volume 1: Matters of Public Interest*. University of Minnesota Press. ISBN: 0-8166-19182.

"The great names of psychology's comparatively recent past are respected mainly as intrepid explorers who came back empty-handed. There is no edifice, just this year's ant hill, most of which will be abandoned and washed away in another season…" p. 7… no sense of cumulative science.

Validity and Utility in I/O Psychology 10th May, 2005 Auckland I/O SIG and HRINZ

"Psychologists in the American Association for the Advancement of Science have been trying recently to get Science to publish a psychological article now and then. The editors reply that they get lots of submissions from psychologists but they just are not as interesting as all the good stuff they keep getting from the biochemists, the space scientists, the astronomers, and the geneticists. Moreover, Science, like its British counterpart, Nature, is a relatively fast publication journal where hot, new findings are published, findings that are of general interest and that other workers in the field will want to know about promptly. But psychologists seldom have anything to show and tell that other psychologists need to know about promptly. We are each working in a different part of the forest, we are not worried that someone else will publish first. and we do not need to know what others have found because ours is not a vertical enterprise, building on what has been discovered previously."

Is the current paradigm slowly dissolving?

VALIDITY

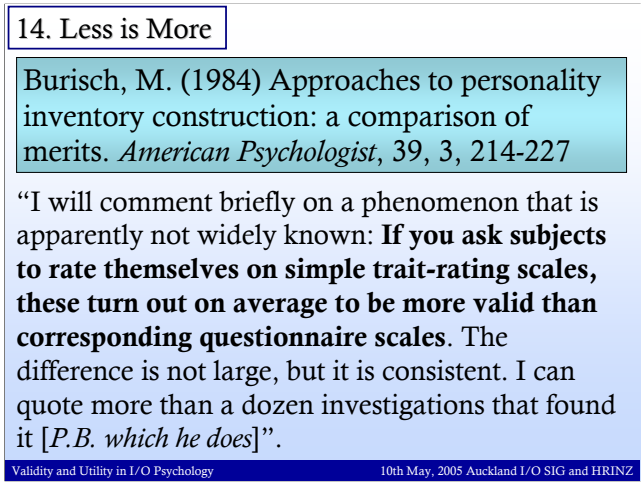
14. Less is More

Burisch, M. (1984) You don't always get what you pay for: measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality*, 18, 81-98.

p. 96 -"The fascination of glamorous data processing devices may indeed have blinded us to some important aspects of reality".
p. 97 -"Given the coarse grain of the language we converse in, we should stop expecting major breakthroughs from polishing verbal instruments."

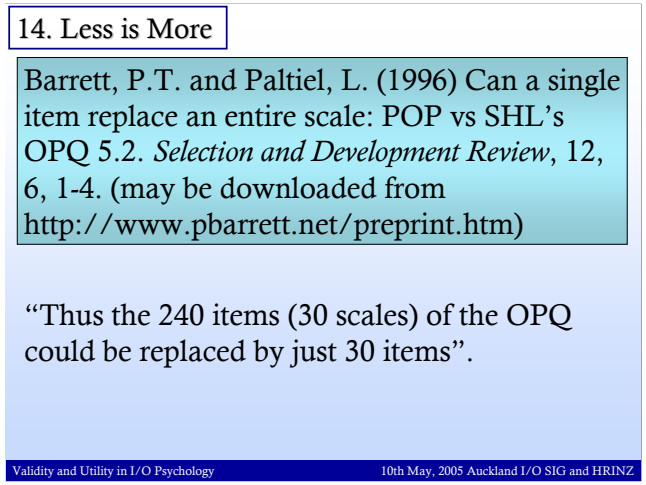
Abstract

In three studies subjects' depressiveness was assessed by a variety of instruments. Questionnaire scales were either comparatively short or long and either fairly simple, content oriented, and undisguised, or sophisticated in the sense of reflecting psychodynamic theorizing or elaborate multivariate approaches to scale construction. Simple self ratings [P.B. single rating items for a whole scale] were also obtained. Results showed that a) short scales were as valid on the average as long scales in all three studies, even though some of the short scales were merely subsets of the long scales; b) simple scales were as valid as sophisticated sales in all three studies; and c) self-rating scales were as valid as questionnaire scales in two studies, but not in the third. The discussion focuses on certain unrealistic assumptions of the Spearman-Brown formula and on the notion of personality assessment as a noise-afflicted communication process.



Abstract

The three major approaches to personality scale construction, the external, the inductive, and the deductive strategies are discussed and their rationales compared. It is suggested that all scales should possess validity, communicability, and economy. The relative importance of these characteristics, however, varies with the purpose for which the instrument is being constructed. A review of more than a dozen comparative studies revealed no consistent superiority of any strategy in terms of validity or predictive effectiveness. But, deductive scales normally communicate information more directly to an assessor, and they are definitely more economical to build and administer. Thus, wherever there is a genuine choice, the simple deductive approach is recommended. Furthermore, self-rating scales narrowly but consistently outdo questionnaire scales in terms of validity and are clearly superior in terms of communicability and economy. **There may not be many situations in which the widespread preference for questionnaires is justified.** It is concluded that the more commonsensical approaches to personality measurement have a lot to offer.



This paper can be read in conjunction with the recent presentation on Single Item Psychometrics and the BPS Occupational Psychology presentation of 1995, (links on the web-page entry as given above to downloadable presentations) which provides even more detail about the POP and OPQ scales, along with the examination of the 16PF and 16PF5 scales (in terms of item homogeneity). The abstract begins ... The Occupational Personality Questionnaire (Concept 5.2) (Saville et al. 1993) contains 248 items measuring 31 scales. Each scale has 8 items. Responses to each item are on a normative 5 point rating scale. Reliability coefficients (alpha) range from 0.57 to 0.88, with a median alpha of 0.75. Since alpha is known to depend on scale length as well as internal consistency, scales of just 8 items may achieve high levels of reliability (greater than 0.7) due to item redundancy e.g. where the items within a scale are simply reworded counterparts of one another. Instead of measuring a broad dimension of behaviour, it is possible that just one rather specific behavioural item is being assessed – using 8 very similar items to achieve this. If this is the case with the Occupational Personality Questionnaire (OPQ), then it should be possible to replace each of the 30 personality scales (excluding the social desirability 'validity' scale) with a single composite item that captures the essential meaning of the scale and its constituent items. **Thus the 240 items (30 scales) of the OPQ could be replaced by just 30 items.**

14. Less is More

Barrett, P.T. (forthcoming) Graphical Profiler Assessment – time to bid questionnaires farewell. A detailed presentation on the technology may be downloaded from: www.pbarrett.net/nz_psych_conference_2003.htm#gpa

The technology powered the Mariner7 and now StaffCV Preference Profiler commercial products, and is now being targeted at personality assessment. Research into this kind of assessment is also being undertaken simultaneously in the US by Prof. James Grice and his colleagues.

One hundred university students were administered a 106-item questionnaire that assessed 10 of the 45 facets of Goldberg's AB5C Five Factor Model Personality Questionnaire. Each student also completed a new prototype of a computer-administered personality assessment that utilised a one-dimensional graphical profiler methodology pioneered by the first author. The 10 facets and 106 questionnaire items were reduced to just 10 single rating statements, with responses made using positioning by a computer mouse of each facet name onto a non-quantitative rating scale bounded by two phrases "Most Like Me" and "Least Like Me". Each participant was also asked which method of assessment they preferred, and which one seemed to allow them to best represent their personality via self-report. Scores acquired from both methods of assessment were compared to one another for direct equivalence, along with analyses that examined the participants' use of the non-quantitative rating scale. Results indicated non-equivalence of assessment method scores, but the majority of participants rated the profiler as the optimal method by which they felt they could describe their personality. An unusual research question for personality psychometrics has now been raised …"which is the most accurate method of assessment of an individual's personality characteristics?"

15. Construct Validity

McGrath, R. (in press – will be published September 2005) Conceptual Complexity and Construct Validity. *Journal of Personality Assessment* – with three commentaries by myself, Michael Maraun, and Jerome Kagan.

"The present article explores the proposition that the *conceptual complexity* of the constructs psychologists choose to measure and the scales they use to measure them has played an important role in the failure to develop more accurate measurement systems".

Abstract

Despite a century of methodological and conceptual advances in the technology of psychosocial measurement, poor correspondence between indicators and the constructs they are intended to represent remains a limiting factor to the accumulation of scientific knowledge. This article presents the hypothesis that longstanding conventions in measurement contribute to the failure to develop optimal criteria. These conventions include the focus on complex over simple constructs, and the use of multi-item measures of disparate content to represent those constructs. A series of arguments is presented suggesting that such a measurement model compromises the potential for developing measures which accurately reflect psychosocial phenomena. Some preliminary suggestions concerning an alternative model that may address this construct validity problem more effectively are offered.

15. Construct Validity

McGrath, R. (in press) …

"One of the basic requirements of science is accurate measurement. Despite a century of effort devoted to improving methods for the creation and evaluation of measurement devices, psychologists generally agree that many if not most of the scales commonly used for the observation of psychosocial events and states provide at best a rough reflection of the constructs they are intended to represent

"Reading research on test validation might lead one to assume that the predictive purposes of tests are more important than their representational purposes. The bulk of the literature dedicated to demonstrating scale validity focuses on scale relationships with expected correlates. In his classic text on assessment, Wiggins (1973) even asserted that "personality assessment has the quite applied aim of generating predictions about certain aspects of behavior that will contribute to decisions concerning the disposition or treatment of individuals" (p. 6), a statement that overlooks the descriptive and model-building aspects of personality research."

15. Construct Validity

Maraun, M.D. (1998) Measurement as a Normative Practice: Implications of Wittgenstein's Philosophy for Measurement in Psychology. *Theory & Psychology*, 8(4), , 435-461.

"Measurement practice in psychology misdiagnoses the nature of measurement, since it is uniformly formulated under the assumption that measurement claims are justified in large part through empirical case-building [aka construct validity]" (p. 436)

Abstract

Recently, a number of prominent measurement specialists (e.g. Cliff, 1992; Schonemann, 1994) have pondered the lack of progress in the development of convincing solutions to the measurement problems of psychology, and have attempted to identify the factors responsible for this lack of progress. They suggest a number of possibilities, including a basic lack of talent in the ranks of the social sciences. It is argued here, however, that the philosophy of Wittgenstein provides an interesting alternative explanation. Specifically, despite their apparent differences, current approaches to the support of psychological measurement claims are unanimous in viewing measurement as chiefly an empirical matter. On Wittgenstein's account, however, this is a mischaracterization of measurement, for, as he argued in elaborate detail, measurement is a normative, rule-guided practice. Hence, empirically based argument is not relevant to the support of measurement claims. If this verdict is correct, it explains not only the failure of measurement theory in psychology, but the much discussed success of measurement in the physical sciences. In this paper, Wittgenstein's characterization of measurement, and its implications for psychology, are discussed.

15. Construct Validity

Maraun, M.D. (1998) Measurement as a Normative Practice: Implications of Wittgenstein's Philosophy for Measurement in Psychology. *Theory & Psychology*, 8(4), , 435-461.

"The problem is that in construct validation theory, ***knowing*** about something is confused with an understanding of the ***meaning*** of the concept that denotes that something….."

Looking at Cronbach and Meehl's (1955) credo of construct validity …

"Scientifically speaking, to 'make clear what something is' means to set forth the laws in which it occurs." … Maraun replies …

"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**." (Maraun … 1998 p. 448)

15. Construct Validity

"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)

15. Construct Validity

Fisher Jr., W. (1997). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement*, (2), 87-113

He argues for the equivalent of a "System International" series of standard unit measures within psychology. This is part of the *normative* process that Maraun postulates – i.e. the rules which are constitutive of the measurement of a construct. The key here is the concept of "metrology".

See also …
Fisher Jr., W. (2005). Daredevil Barnstorming to the Tipping Point: new aspirations for the human sciences. *Journal of Applied Measurement*, 6, 2, 173-179. A very interesting account of what it might take to develop a coherent, organized, and unified measurement system for psychology, which would include metrology (as does physics in countries' National Physical Laboratories around the world).

15. Construct Validity

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004) The concept of validity. *Psychological Review*, 111, 4, 1061-1071.

"Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that **a test is valid if it measures what it purports to measure**.

Abstract

This article advances a simple conception of test validity: A test is valid for measuring an attribute if

(a) the attribute exists and

(b) variations in the attribute causally produce variation in the measurement outcomes.

This conception is shown to diverge from current validity theory in several respects. In particular, the emphasis in the proposed conception is on **ontology, reference, and causality**, whereas current validity theory focuses on **epistemology, meaning, and correlation**. It is argued that the proposed conception is not only simpler but also theoretically superior to the position taken in the existing literature. Further, it has clear theoretical and practical implications for validation research. Most important, validation research must not be directed at the relation between the measured attribute and other attributes but at the processes that convey the effect of the measured attribute on the test scores.

15. Construct Validity

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004) The concept of validity. *Psychological Review*, 111, 4, 1061-1071.

"The argument to be presented is exceedingly simple; so simple, in fact, that it articulates an account of validity that may seem almost trivial. It is as follows. 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".

"We start this article with a request to the reader. Please take a slip of paper and write down your definition of the term *construct validity*. Now, take the classic article of Cronbach and Meehl (1955), who invented the concept, and a more recent authoritative article on validity, for instance that of Messick (1989), and check whether you recognize your definition in these works. You are likely to fail. The odds are that you have written down something like "construct validity is about the question of whether a test measures what it should measure." If you have read the articles in question carefully, you have realized that they do not conceptualize validity like you do. They are not about a property of tests but about a property of test score interpretations. They are not about the simple, factual question of whether a test measures an attribute but about the complex question of whether test score interpretations are consistent with a nomological network involving theoretical and observational terms (Cronbach & Meehl, 1955) or with an even more complicated system of theoretical rationales, empirical data, and social consequences of testing (Messick, 1989)."

So now what?

For the Academic

- Basically ignore or reject most of the above the above and carry on as normal.
- Critically evaluate some of the above evidence and logic - and then decide what this will mean for you personally if you agree with say Michell, and/or Gigerenzer, and/or Borsboom et al.

So now what?

For the I/O Practitioner

• Well, what now for those who sell or use ipsative tests?
• What about those who promulgate "team-working" as optimal for increased productivity and job performance?
• What happens when single-item psychometrics hits the market?
• How will you now speak about Validity and Psychometrics?

• At best, the evidence and arguments above suggest that what I/O psychologists promulgate as "useful", "valid", or "evidence-based", is not as strongly supported by the actual evidence, argument, or fundamental logic as they might have clients believe. Whether this matters to clients is a moot point. If clients decide it does, and if individuals like myself are available to address these issues directly and openly whilst solving their problems, who knows what effect this will have on the I/O profession?

In this regard, the paper by Swets, J.A., Dawes, R.M., & Monahan, J. (2000) Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest*, 1, 1, 1-26 discusses how certain practice outcomes might need to be evaluated. But, sweeping innovations in testing, modeling with situational as well as psychological variables, and a much greater sense of empirical reality regarding the validity of certain I/O interventions will likely cause fundamental changes over time about what is expected from I/O psychologists, and what they can realistically be asked to deliver. But, I don't expect any major changes at all in the short term, and any will depend heavily upon how academics deal over time with the evidence and argument from the various sources quoted here. All I am really saying is that the pressure for change is building – and it looks like a paradigm change – for example, not merely a transition from say Classical Test Theory to IRT – but the entire loss of personality and maybe even intelligence questionnaires and psychometric test theory altogether over time. The new definition of Validity is huge in its implications for the whole of psychology and its measurement of constructs, not just personality and I/O psychology.