Citation: Barrett, P.T. * Prinsloo, M. (2013). Rethinking reliability and validity of psychological measurements. *> 0.05: The Non-Significant Journal of Business and Consumer Psychology*, Autumn 2, 2, 28-55.

Rethinking Reliability and Validity of psychological measurements

Paul Barrett & Maretha Prinsloo

INTRODUCTION

This article draws upon the thinking and writing of Denny Borsboom and others on the demise of construct validity, as well as providing an Appendix defining the constituent properties of quantitative measurement (which makes crystal clear why no psychological attributes can be claimed as varying quantitatively right now; there is simply no evidence to support such an assumption).

What we wanted to show was that it is still possible to establish the reliability and validity of any instrument, process, HR intervention, coaching scheme, or consultancy process, that would satisfy the requirements for a legally defensible evidence-base, even though in many cases what we are dealing with is non-quantitative variation, differing amounts of subjective 'perturbations' on assessment results, and usually imprecise outcomes.

We have attempted to integrate various sources to produce what we think looks like a coherent evidence-base-construction strategy that is sensitive to the reality of how psychological assessments/interventions are actually deployed and used in the workplace.

There are basically two lines of argument put forward:

1. That reliability assessment is not about computing Cronbach's alpha or the variety of ad-hoc factor-loading/latent variable methods proposed as 'measures of reliability'.

2. That assessing the validity of a psychological assessment is far removed from any of the 'usual suspect' definitions of validity that permeate the literature. E.g. construct validity is finished as a viable concept.

PREAMBLE

How do we assess the reliability of a measure which, by its very nature, precludes re-testing within a period of time where familiarity of what was tested previously will distort future scores on the measure? When we investigate validity, we have two questions to answer; the first is concerned entirely with measurement. The second question is concerned with meaning:

1. Does the test measure what it claims to measure?
2. Does the test score show the expected relationships with other theoretically-relevant scores, behaviours, and outcomes?

But, from consideration of an alternative perspective on validity, another simple question arises for which an answer can be sought:

Do clients find substantive value in using psychometric tests?

Reliability

Consider the fundamental definition of reliability found within any science as well as in everyday usage:

The repeatability problem for psychological assessment is noted by Guttman as far back as 1945 in his article entitled: "A basis for analysing test-retest reliability".

"The problem of reliability is of course not peculiar to psychology or sociology, but pervades all the sciences. In dealing with empirical data in any field, the question should be raised: if the experiment were to be repeated, how much variation would there be in the results?"
p.256

And further on page 257:
"A major emphasis of this paper is that the reliability coefficient cannot in general be estimated from but a single trial-that items do not replace trials. If two trials are experimentally independent, then we show that the correlation between two trials is, with probability of unity, equal to the reliability coefficient.

As is well known, there may be great practical difficulties in making two independent trials; therefore our principal focus is on *what information can be obtained from a single trial.* We find that *lower bounds* to the reliability coefficient can be computed from a single trial. Six different lower bounds are derived, appropriate for different situations. Several of these bounds are as easy as or easier to compute than are conventional formulas, and all of the bounds assume less than do conventional formulas.

To prove that bounds can be computed from a single trial, we use essentially one basic assumption: that the errors of observation are independent between items and between persons over the *universe of trials.* In the conventional approach, independence is taken over *persons* rather than trials, and the problem of observation from a single trial is not explicitly analysed."

Therein lies the convenient concept deployed by psychometricians – a hypothetical *universe of trials*.

By making an assumption based upon statistical sampling theory, and invoking the concept of a universe of items which 'measure' a single attribute, from which a random sample has been drawn by the investigator *(the items in any particular test),*it is possible to generate a variety of bounds for reliability, which is exactly what Guttman achieved in his article.

Reliability is the extent to which a second, third and onwards observations of an event, measure, rating, or occurrence, deviates from the first or proceeding observations. If they are exactly the same, there is perfect reliability. In engineering terms, reliability is referred to as repeatability.

Cronbach (1951) extended Guttman's work and introduced the now famous Cronbach alpha coefficient. This made reliability assessment a routine feature of analysis, augmenting the simple definition of reliability as repeatability by using that key assumption of an investigator sampling items from a 'universe' of items, with the additional proposition stating that individuals can be

administered a multitude of parallel tests drawn from that universe. From this assumption-laden test-theory definition other kinds of reliability coefficients were constructed, such as omega, factor validity, and the complex indices that have been constructed recently in Structural Equation Modelling *(Cortina (1991), Schmitt (1996), Green and Yang (2009) provide good overviews)*.

However, all these indices depend for their validity upon assumptions made about test scores as quantities, hypothetical true scores, and hypothetical item universes. The real bottom line to reliability stays the same, regardless of what psychometricians might wish otherwise.

It's about answering *"what happens with an individual's scores 2nd or 3rd time around on this test"*. Not a population of tests, not a population of people, but **this person, this test, right now**.

Ordinarily retest reliability and a discrepancy-score work-up would answer this question nicely, except that retest reliability estimation now includes three sources of 'perturbation':

- Non-systematic random error associated with the internal integrity of the test itself.
- Systematic attribute variation on the attribute over periods of time within and extending across the retest duration.
- Memory of previous responses artificially causing consistency in 2nd occasion response patterns.

The end result of a retest analysis would nevertheless be an indication of reliability, but the causes of any substantive unreliability are not able to be disentangled from each other except by further careful empirical investigation.

Validity

When we concern ourselves with validity, we have two questions to answer; the first is concerned entirely with measurement, the second with meaning:

- Does the test measure what it claims to measure?
- Does the test score show the expected relationships with other theoretically relevant scores, behaviours, and outcomes?

Many are confused about how to define and investigate the validity of a psychological assessment, confounding issues of measurement with those of the perceived/proposed consequences of assessed attribute variations.

From a **measurement perspective**, the validity of any measurement process is concerned solely with the accuracy, reliability, and precision, of any proposed measurement scheme reflecting the variations of magnitude of an attribute which is claimed to be measured by the test or magnitude evaluation process.

In terms of **consequences**, whether or not attribute variation should relate to, be causal for, or predict other phenomena is a matter for a different kind of theory-guided empirical investigation. Many psychologists mistakenly view this as the process of developing assessment validity through

building what Paul Meehl referred to as a 'nomological net' or a complex web of other-attribute relationships.

This is quite wrong. A detailed explanation of why is provided in two chapters in the book edited by Lissitz (2009) entitled: “The Concept of Validity: Revisions, New Directions, and Applications”.

Point 1: Chapter 6, Michell, J. (2009a). Invalidity in Validity.

“Abstract: The concept of test validity was proposed in 1921. It helped allay doubts about whether tests really measure anything. To say that the issue of a test's validity is that of whether it measures what it is supposed to measure already presumes, first, that the test measures something and, second, that whatever it is supposed to assess can be measured. An attribute is measurable if and only if it possesses both ordinal and additive structure. Since there is no hard evidence that the attributes that testers aspire to measure are additively structured, the presumptions underlying the concept of validity are invalidly endorsed. As directly experienced, these attributes are ordinal and non-quantitative. The invalidity in validity is that of feigning knowledge where ignorance obtains. “

Point 2: Chapter 7, Borsboom, D., Cramer, A.O.J., Kievit, R.A., Scholten, A.Z., & Franic, S. (2009). The end of construct validity.

“Abstract: Construct validity theory holds that validity is a property of test score interpretations in terms of constructs that reflect the strength of the evidence for these interpretations. In this paper, we argue that this view has absurd consequences. For instance, following construct validity theory, test score interpretations that deny that anything is measured by a test may themselves have a high degree of construct validity. In addition, construct validity theory implies that now defunct test score interpretations, like those attached to phlogiston measures in the 17th century, 'were valid' at the time but 'became invalid' when the theory of phlogiston was refuted. We propose an alternative view that holds that (a) validity is a property of measurement instruments that (b) codes whether these instruments are sensitive to variation in a targeted attribute. This theory avoids the absurdities of construct validity theory, and is broadly consistent with the view, commonly held by working researchers and textbook writers but not construct validity theorists, that a test is valid if it measures what it should measure. Finally, we discuss some pressing problems in psychological measurement that are salient within our conceptualization, and argue that the time has come to face them.”

Does the test measure what it claims to measure?

First and foremost, we need to be very specific about what is meant by that word ‘measurement’. Within the particular theory that underpins all natural science measurement, the word has a very specific, and a very restrictive meaning. Michell (1999).

Clearly, given the constitutive properties of measurement that characterize the SI units measures of physics (http://physics.nist.gov/cuu/Units/units.html), the constructs, attributes, processes, and preferences reported upon within the majority of psychometric data are not quantities; this includes ability, IQ, personality, learning potential, motivation, spiral dynamics, Jacques stratified systems,

and all latent variables which are simply declared to vary quantitatively by psychologists, *in absentia* of any evidence supporting that knowledge-claim.

It is important readers understand that it does not automatically follow that when the assignment of numbers to represent attribute magnitudes is undertaken, the attributes themselves do actually vary as quantities. For example, do the attributes of Conscientiousness or Agreeableness vary additively within any individual, as does say mass or thermodynamic temperature?

That question can only be answered by careful empirical experimentation which might reveal the quantitative structure of the attribute.

Does the test score show the expected relationships with other theoretically relevant scores, behaviours, and outcomes?

The conventional route of concurrent and predictive validity attempts to construct a nomological network. In this respect, the 'usual' approach to validation is that which any high-quality professional test publisher might adopt and report upon. In a sense, the 'usual suspect' boxes of test validation have been ticked.

But so much subjective interpretation of test scores takes place by users that the very nature of what constitutes evidence of validity is rendered problematic, suggesting an altogether different investigative/evaluative strategy.

As we have said many times before, the problem is that the evidence for the validity of test scores is rendered *uncertain* in actual practice, because psychologists, HR, those who actually use the tests, do not treat the scores as measurements *(as we do say a measurement of length or mass)*, but as *indicative indices* to be more, or less, subjectively interpreted in the context of a body of other information.

Not only that, in many cases the user never sees a test score, only a transformed version expressed as a 'normative' score.

The degree to which test data and test reports invite differences in test interpretation varies greatly. However, for any psychological test, it is only in those cases where the test score itself is used as the sole basis for a decision (a cut-off score/threshold range) that the evidence for effects reported in test publisher manuals or academic publications are likely to be seen in real-world applications.

Let us put this issue into a real-world context:
The following brief details of a case-study describes an assessment used to predict the risk of violent recidivism in offenders seeking parole; the risks involved in transferring forensic-psychiatric mentally-disordered offenders to lesser-security institutions; and the risk-profiling in general of offenders who may be approaching their applicable release date from incarceration.

Harris, Rice, and Quinsey (1993) collected personal, clinical, and offence-related information in this regard to form a scale of the predictors of the risk of violent recidivism; assigned weights to these

predictors (reflecting their importance to the prediction); and subsequently develop a cumulative scale of risk (the VRAG). More details are provided in Appendix 2.

Initial empirical exploration of the candidate information most predictive, in combination, of violent recidivism over a fixed period, produced 12 predictor attributes. The first 11 predictors are a mixture of bio data and clinical diagnoses (such as a history of alcohol problems and age index offence). Predictor #12 is a psychological assessment, a rating checklist whose ratings on affective and behavioural attributes are provided either by trained assessors from patient/offender records or by offender interview with a trained forensic clinical psychologist or forensic nurse.

The process of forming the risk assessment scale was taken from the Webster et al (1994) test manual, pp. 33-34, according to which the sample was divided into subgroups of varying risk levels. A risk propensity graph with test scores of 1 to 9 was devised.

The information imparted by such a graph is straightforward. All the information required to interpret the score is given by the probability of occurrence of the criterion outcome. Normative scores would impart zero benefit.

The subjectivity of interpretation of any new individual's scores is encountered when considering two questions:

1. Could this new individual be considered to share sufficient similarity with the calibration sample group such that the predictions made using the sample data could be reasonably inferred to apply to that individual?
2. What level of risk constitutes an unacceptable level?

There is no requirement or need for a 'narrative report', some kind of subjective interpretation of "risk personality", or trying to interpret the what the correlation of 0.45 between VRAG scores and recidivist risk means with regard to an offender's or patient's risk. Note also that scoring is *algorithmic*, carefully designed and calibrated against the criterion of interest.

The VRAG does not measure a quantity. But, by the very nature of its design and calibration, the 'indicative indices' {1..9} to which we referred earlier produce a straightforward assessment of risk, that of the probability of recidivist outcome over a 7 year period post-release. Any other *interpretation* of the 'scores' is entirely subjective.

The score is interpreted by a trained practitioner who may also prepare a final report which integrates all the information presented in the VRAG with additional interpretations of data from other assessment tools.

The validity of the judgements made and decisions taken about the test-taker thus depends on both the test results as well as the interpretations of the report. In fact, interpretations may well depart significantly from the content of the VRAG data, *reflecting bias, a lack of understanding, and a faulty integration of external information.*

The issue of the validity of interpretation actually plays a role in most professional settings where theoretical guidelines are interpreted and applied by experts. It also pertains to all psychometric practices where practitioners make use of self-report personality questionnaire test reports. We have lost count of the numbers of times I've seen HR executives, consultants, and psychologists treat computer-generated narrative reports as *indicative* or *suggestive* rather than prescriptive assessment tools.

Very rarely are cut-scores employed, and usually only as a pre-screen to whittle down candidate numbers to a more manageable subset. From then on, the interpretation skills of whoever is charged to form judgements and make decisions comes into play, with the test scores relegated to the status of *'one of many sources of information to be considered and integrated*'.

The reality is that for the vast majority of psychological assessments, the validity requirement for an evidence-base includes the validity of decisions made by the interpreters of test scores *(consultants, HR, psychologists, business-school lecturers etc.)*, as well as the test scores. That complicates matters.

The complication is that we have to take into account the relative skills and attributes of interpreters of test scores as well as the test scores themselves, when looking at the reported utility of an assessment by users. Accordingly, we can sidestep the measurement/assessment related/psychometrics issues by asking a simple question:

"Do clients find substantive value in using an assessment?"

Let's look more closely at the reasoning.
If an assessment technique is 'unreliable' in the sense that its assessment of attributes is near random *(the report content reflects near-random individual designations, scores, and preferences)*, clients would soon discover that what the report says about an individual does not accord with other known information about that person.

Furthermore, it would show no systematic relationships with theoretically-relevant attributes. However, let's assume the assessment is reliable. It may simply be assessing attributes which have no bearing on, or relevance to, what clients wish to know about an individual in order to make certain inferences/judgements/decisions about that individual.

The end result in either case is that clients using the assessment would find that it has provided no tangible utility at all. Decisions that were made on the basis of the assessment would be seen to be inaccurate; heavy costs would be incurred through 'failed' selection, promotion, coaching, or other kinds of "workforce intervention" decisions. Consultants would very quickly avoid using it as it would be reflecting poorly on their own work/judgements, and negatively impacting their income, (likewise for those using it in HR).

The immediate criticism which can be applied to this line of reasoning is that many practitioners involved with employee coaching, development, recruitment, selection, team-functioning,

leadership, and career-path trajectory planning will attest to the utility of graphology, DISC-based assessments, type-based assessments such as the MBTI, and various EQ/EI assessments.

But *all* of these have been shown to be of near-zero to dubious 'validity' within many peer-reviewed scientific journal publications.

But, can it really be the case that for a worldwide best-selling assessment like the MBTI, for which no replicable empirical evidence exists for discrete 'type' separation as claimed by its protagonists, users are simply misguided 'believers' who assume utility where none is manifest?

On the contrary, the assessment must be demonstrating observable utility, given the continued use and purchase of the product, even though the academically-generated evidence suggests it should not be doing so.

The reasons why this paradox exists are complex, and would justify the writing of another article devoted to just this issue. However, one major consideration is how psychological assessments like these are used in the workplace. Their results are heavily interpreted by users, forming the basis for deep psychological reasoning about an individual, their interpersonal interactions, and a level of debate and discussion around a report that swamps the information imparted by any single score, fragment of hand-written text, or classification.

That *practitioner-led* interpretation of results, and the associated interactivity, human synthesis and integration of information around the test results is the systematic factor which is missing from all peer-reviewed research which is focused solely on the scores or classifications.

Furthermore, the practitioner factor can be expected to attenuate aggregate-effect validity studies because practitioners within specific organizational implementations may produce positive and negative effects, limiting the magnitude of what would otherwise be systematic relationships between assessment results and outcomes. Clearly, finding evidence for any assessment tool is not simply a matter of correlating a few scores with some supervisor ratings.

Two articles bear upon this issue; the most recent is authored by Bornstein (2012), entitled: "*Rorschach score validation as a model for 21st-century personality assessment*", concerning the strategy for validating the objective scoring systems for the Rorschach Inkblot Test:

"**Abstract**: Recent conceptual and methodological innovations have led to new strategies for documenting the construct validity of test scores, including performance-based test scores. These strategies have the potential to generate more definitive evidence regarding the validity of scores derived from the Rorschach Inkblot Method (RIM) and help resolve some long-standing controversies regarding the clinical utility of the Rorschach. After discussing the unique challenges in studying the Rorschach and why research in this area is important given current trends in scientific and applied psychology, I offer 3 overarching principles to maximize the construct validity of RIM scores, arguing that (a) the method that provides RIM validation measures plays a key role in generating outcome predictions; (b) RIM variables should be linked with findings from neighbouring subfields; and (c) rigorous RIM score validation includes both process-focused and outcome-focused

assessments. I describe a 4-step strategy for optimal RIM score derivation (formulating hypotheses, delineating process links, generating outcome predictions, and establishing limiting conditions); and a 4-component template for RIM score validation (establishing basic psychometrics, documenting outcome-focused validity, assessing process-focused validity, and integrating outcome- and process-focused validity data). **The proposed framework not only has the potential to enhance the validity and utility of the RIM, but might ultimately enable the RIM to become a model of test score validation for 21st-century personality assessment**. " (p. 26).

The second article is authored by Nutt (1999), entitled: "*Surprising but true: Half the decisions in organizations fail*";

"**Abstract**: Half the decisions in organizations fail. Studies of 356 decisions in medium to large organizations in the U.S. and Canada reveal that these failures can be traced to managers who impose solutions, limit the search for alternatives, and use power to implement their plans. Managers who make the need for action clear at the outset, set objectives, carry out an unrestricted search for solutions, and get key people to participate are more apt to be successful. Tactics prone to fail were used in two of every three decisions that were studied." (p. 75)

He outlines the research strategy in a paragraph further on the same page:
"To find out why decisions go wrong, I began my research by collecting real decisions in real organizations, made by real people. Getting close to the action uncovered tactics and allowed me to see a decision's result and its consequences. Connecting outcomes to tactics provided a telling appraisal of the effectiveness of the tactics employed by managers. " (p. 75)

This article is the exemplar of the "shoe-leather" research strategy proposed by the famous statistician, David Freedman (1991), in an article entitled "*Statistical models and shoe leather*":

"**Abstract:** Regression models have been used in the social sciences at least since 1899, when Yule published a paper on the causes of pauperism. Regression models are now used to make causal arguments in a wide variety of applications, and it is perhaps time to evaluate the results. No definitive answers can be given, but this paper takes a rather negative view. Snow's work on cholera is presented as a success story for scientific reasoning based on non-experimental data. Failure stories are also discussed, and comparisons may provide some insight. In particular, this paper suggests that **statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings**" (p. 291)

Having "ticked" the conventional boxes for assessment validation, we are now taking validation into new but more meaningful territory. Validation involves asking the practitioners and users of tests whether they have found utility in using them. We may partition users into various categories but the bottom-line question is simple:

"Has your use of the test added value to your work and/or your organization?"

We are not interested in detail, simply whether or not users have found the test in question worthwhile. It's a crude index of real-world validity, but if the answer is positive, it amounts to a clear indicator of perceived utility.

We are designing an actuarial approach to acquiring evidence of specific effects. That is, we utilise the data acquisition strategy of the VRAG, the shoe-leather from Freedman, and the 'real-world' focus of Nutt. In essence, we seek to generate an evidence-base of claim versus outcome for practitioners, consultants, and those in organizations who made certain decisions about individuals based upon CPP indications/recommendations.

Some examples:

- Incumbent employees being selected for managerial/supervisor/leadership roles. How many actually succeeded over time in those roles?
- Incumbent employees identified as possessing "potential", and selected for leadership development/training. How many eventually demonstrate that potential?
- Candidates selected for particular job-roles. How many succeeded in those roles?
- Senior executive recommendation/appointments. How many proved to be successful.
- Candidates/incumbents identified with a problem-solving preference or preferred work environment. How many showed the expected consequences of those preferences in their work-roles and performance outcomes?

The evidential import of such data cannot be underestimated. Instead of the largely abstract validity coefficients put forward by so many test publishers which mostly relate to broad, generic outcomes, this evidence speaks directly to the frequency of very specific outcomes which can be directly associated with the use of a specific test. And because trajectories are person-specific, subsequent aggregation of complex outcome trajectories is optional and not a mandatory feature of standard *validity-generalization* statistical strategies.

In Conclusion

It is hoped that the reader now has an appreciation of an alternative approach towards test reliability and validity.

Take-aways

- Reliability and validity for a performance-based test is a function both of interpreter and objective test scores and designations.

- Test reliability and validity should answer that simple question "*does it do what it says on the box?*" with clear answers and appropriate evidence.

- Abstract psychometric parameters and data models which are predicated upon untested assumptions of attribute quantitative variation is not an appropriate strategy for answering such a practical question.

References

Bornstein, R.F. (2012). Rorschach score validation as a model for 21st-century personality assessment. *Journal of Personality Assessment*, 94, 1, 26-38.

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Scholten, A.Z., & Franic, S. (2009). The end of construct validity. In Lissitz, R.W. (Eds.). *The Concept of Validity: Revisions, New Directions, and Applications* (Chapter 7, pp. 135-170). Charlotte: Information Age Publishing.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 1, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 3, 297-334.

Freedman, D.A. (1991). Statistical models and shoe leather. *Sociological Methodology*, 21, 1, 291-313.

Freedman, D.A., Collier, D., Sekhon, J.S., & Stark, P.B (Eds.), (2009). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge UK: Cambridge University Press.

Green, S.B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74, 1, 155-167.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 4, 255-282.

Lissitz, R.W. (Ed.). (2009). *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte: Information Age Publishing.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. New York: Lawrence Erlbaum.

Michell, J. (1994) Numbers as quantitative relations and the traditional theory of measurement. *British Journal for the Philosophy of Science*, 45, 389-406.

Michell, J. (1997). Quantitative science and the definition of measurement in Psychology. *British Journal of Psychology*, 88, 3, 355-383.

Michell, J. (1999). *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge University Press. ISBN: 0-521-62120-8.

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

Michell, J. (2009a). Invalidity in Validity. In Lissitz, R.W. (Eds.), *The Concept of Validity: Revisions, New Directions,and Applications* (Chapter 6, pp. 111-133). Charlotte: Information Age Publishing.

Michell. J. (2009b). The psychometricians' fallacy: Too clever by half? B*ritish Journal of Mathematical and Statistical Psychology*, 62, 1, 41-55.

Michell, J. (2012a). "The constantly recurring argument": Inferring quantity from order. *Theory and Psychology,*22, 3, 255-271.

Michell, J. (2012b). Alfred Binet and the concept of heterogeneous orders. Download link: http://www.frontiersin.org/quantitative_psychology_and_measurement/10.3389/fpsyg.2012.00261/abstract *Frontiers in Quantitative Psychology and Measurement*, 3, 261, 1-8.

Michell, J., & Ernst, C. (1996). The Axioms of Quantity and the Theory of Measurement: Translated from Part I of Otto Hölder's German Text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, 40, 2, 235-252.

Michell, J., & Ernst, C. (1997). The Axioms of Quantity and the Theory of Measurement Translated from Part II of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, 41, 3, 345-356.

Nutt, P.C. (1999). Surprising but true: Half the decisions in organizations fail. *Academy of Management Executive*, 13, 4, 75-90.

Saint-Mont, U. (2012). What measurement is all about. *Theory and Psychology*, 22, 4, 467-485.

Schmitt, N. (1996). Uses and Abuses of Coefficient Alpha. *Psychological Assessment*, 8, 4, 350-353.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science: Part A*, 42, 4, 509-524.

Sijtsma, K. (2009a). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74,1, 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, 74, 1, 169-173.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory and Psychology*, 22, 6, 786-809.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.). *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds), *Measurement: Definitions and Theories*, pp. 18-63. New York: Wiley

The Violence Risk Assessment Guide (**VRAG:** Webster, C.D., Harris, G.T., Rice, M.E., Cormier, C., & Quinsey, V.L. (1994) *The Violence Prediction Scheme: Assessing Dangerousness in High Risk Men.* University of Toronto, Centre of Criminology).