Concept Whitepaper

1

June, 2013

Paul Barrett[1] & **Maretha Prinsloo**[2]

[1]Chief Research Scientist, Cognadev UK.

[2]Director at Magellan SA, and Cognadev UK.

Investigating the reliability and validity of the Cognitive Process Profile (CPP)

How do we assess the reliability of an assessment which, by its very nature, precludes re-assessment within a period of time where familiarity of what was undertaken previously will distort future performance on the assessment? This is the conundrum facing investigation of reliability of the CPP.

When we investigate validity, we have two questions to answer, the first is concerned entirely with measurement, the second 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 the CPP?

Reliability

The fundamental concept behind the entire CPP assessment is novelty; we are assessing an individual's cognitive processes and characteristic features when faced with an entirely novel cognitive task.

The use of novel and unfamiliar information for purposes of cognitive assessment has a number of advantages including triggering metacognitive involvement; reducing the contaminating impact of acquired and content related processing habits; and in being unfamiliar to everyone, it has a levelling effect in cross-cultural contexts. Unfamiliar and fuzzy or vague task content also requires the use of judgement and intuition, which is of particular relevance in the everyday work environment.

If you remove that novelty, you destroy the validity and usefulness of the assessment. So, how do we assess reliability of the CPP?

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

> 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.

The repeatability problem for psychological assessment is noted by Guttman as far back as 1945 in his article entitled: "A basis for analyzing test-retest reliability"… p. 256:

> "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?"

And on page 257:

> "(3) 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.
>
> (4) 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.
>
> (5) 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 observability from a single trial is not explicitly analyzed."

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.

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 Modeling *(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':

1. non-systematic random error associated with the internal integrity of the test itself.
2. systematic attribute variation on the attribute over periods of time within and extending across the retest duration.
3. 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.

However, with the CPP, we know 3 will invalidate the raison d'être of the assessment, which is to investigate how a person's cognitive processes and preferences engage with a novel series of tasks. Once an individual has completed the assessment, the novelty is lost. It can perhaps take years for the memory of the task to fade, which then brings 2 back into focus.

Ultimately, I don't think the reliability of the CPP can be assessed, except indirectly by the assessment of its validity. This is not a problem with the test per se, but is a property of a class of such tests which would seek to assess *'response patterns to novel stimuli'*.

If a parallel version of the CPP could be constructed, then yes, parallel form reliability could be computed. But, the CPP is not a simple assessment to construct *(unlike so many self-report personality or ability item questionnaires that can be so constructed)*. It assesses dynamic learning and cognitive processing using a particular kind of task structure.

The Learning Orientation Index (LOI), which is based on the same methodological approach and assessment techniques as the CPP, but which contains completely different task contents, currently offers as yet unexplored possibilities in this regard. Qualitative and ad hoc studies have already indicated a fair degree of alignment between the CPP and LOI results, particularly on the constructs of cognitive style, "left" versus "right brain" preferences and units of information, but more evidence on this is required.

However, what we do know is that if a test is not reliable, it cannot be valid. That is the simple consequence of 1, and not knowing the conditions causing occasion-to-occasion variability in responses for 2.

But if it is reliable, then it may be valid *(if it assesses what it purports to assess)*. That logic allows us to indirectly infer the CPP reliability from its validity investigations. But it is a logical inference following from the fact that if the assessment claims about an individual made by the CPP are shown to be 'as stated' or eventuate over time, then we know the CPP must be reliable. However, we cannot quantify that 'unreliability' except with reference to the unreliability of the 'knowledge claims' made by the test about an individual, which would also be interpreted as evidence for its validity. For the CPP, reliability and validity are necessarily intertwined; not as a design error but as feature of the particular kind of assessment methodology which the CPP embodies.

Validity

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

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 a 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".

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."

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 (a) validity is a property of test score interpretations in terms of constructs that (b) reflects 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."

With regard to the CPP and addressing the question 1 ***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. Here, I'm going to borrow the definitions provided by Michell (1999).

Measurement	The discovery or estimation of the *ratio* of a *magnitude* of a *quantity* to a *unit* of the same quantity.
Quantity	An attribute possessing ordinal and additive structure. For example, length is a quantity because lengths are ordered according to their magnitude and each specific length is constituted *additively* of other specific lengths.
Magnitude	A specific level of a quantitative attribute (or quantity). For example each specific length that any object might have is a magnitude of the attribute, length.
Unit	A specific magnitude of a quantity relative to which measurements are made. For example, a standard unit of mass is the kilogram; a standard unit of time is seconds.
Ratio	The magnitude of one level of a quantitative attribute to another of the same attribute. If the quantitative attribute is continuous then the ratios are real numbers and they are always relative to a specific relation of additivity.

Technical Appendix 1 provides the technical background to the conditions of quantity, moving from class (type), order, through additivity. It is the additivity conditions which define quantity. Within classical measurement theory, the word *measurement* is confined solely to the measurement of a quantity, which requires a standard unit of a quantity to be specified, against which ratios of magnitudes of the same quantity may be expressed as 'measurements'.

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 CPP assessment data are not quantities; there is no evidence that any of the CPP attributes vary as quantities. But neither is there any evidence that any psychological attribute varies as a quantity. 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.

I don't want to dwell on this issue, but 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.

However, the assessment of counts (frequencies of occurrence), ordered scores, and classes (types) are all properties of the CPP. We may even retain the use of quantitative arithmetic operations for convenience *(while implicitly acknowledging the limitations of the assumed precision)*. These properties set it apart from other tests of learning/performance potential which do not form their judgments based upon the autonomous acquisition and empirical, objectively applied, rule-based scoring of an individual's performance on one or more relevant tasks. This objectivity in scoring enables us to seek comparability with other assessments of the same attributes, in order to demonstrate substantive equivalence between same-attribute assessments.

However, this is complicated by the fact that many of the CPP attributes are unique, in that, except for the LOI which is currently still in the process of being validated, there are no readily available alternative assessments which might be considered from first principles as sufficiently 'equivalent' to a CPP assessed attribute. Although as reported in the technical documentation for the CPP, some work has been undertaken to show relationships between theoretically-relevant attributes, the evidence-base addressing the *'does this test assess what it purports to assess'* question needs a different kind of experimental work, using a variety of calibration tasks designed specifically to assess one or more of the specific preferences or processes currently claimed to be assessed by the CPP. This is an ongoing R&D strategy within Cognadev.

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

To date, the CPP validation activity has followed the conventional route of concurrent and predictive validity workups, in an attempt to construct a nomological network. In this respect the approach to validation is that which any high-quality professional test publisher might adopt and report upon. In a sense, the 'usual suspect' boxes have been ticked, inasmuch as SHL, Kenexa, Psytech, TalentQ, Saville Consulting, and others tick the boxes of 'psychometrics validation'. But so much subjective interpretation of assessment 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 I 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. For example, the CPP report is based upon acquired performance data, not self-report items. The attributes it reports upon directly reflect features of performance which are themselves constructed from empirical counts/frequencies of certain kinds of cognitive-behavioural occurrences, preference choices, and speed of information intake and processing.

But for any psychological test, it is only in those cases where the test score itself is used as the sole basis for a

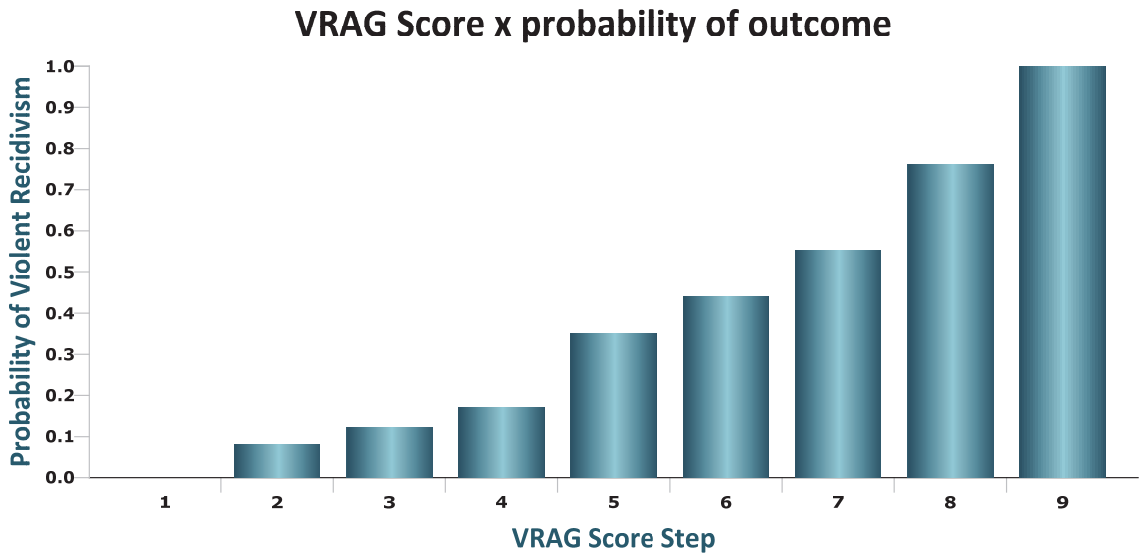
decision (a cut-score/threshold range) that the evidence for effects reported in test publisher manuals or academic publications are likely to be seen in real-world applications.

Let me put this issue into a real-world context.

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) ... see also Rice (1997).

This monograph describes an assessment used to predict the risk of violent recidivism in offenders seeking parole; the risks involved in transfering 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; assign 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 biodata 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 raters 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.



VRAG Score	1	2	3	4	5	6	7	8	9
VRAG Prob.	0.0	0.08	0.12	0.17	0.35	0.44	0.55	0.76	1.0

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 I 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.

Imagine for one moment if that VRAG graph above referred to a learning/information processing preference and career outcome probability. Would we seek to interpret such a relationship at all except from the more scientific perspective of understanding why we are able to observe it?

With specific regard to the CPP, the report generated by the scoring software is intepreted by a trained practitioner who may also prepare a final report which integrates all the information presented in the CPP report, and even integrate the CPP results with that of other assessment tools.

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

The potential impact of fators that may affect the accuracy of the interpretation (such as bias and misunderstanding) have been addressed in a number of ways to ensure a standardised application of the CPP results. Such measures include: product related training and accreditation of professionally registered practitioners; repeated refresher courses; narrative explanations in the CPP report of the meaning and application of the constructs measured; the creation of customised, algorithmically based reports by which CPP constructs are linked to company or job specific competency requirements; and the use of the Contextualised Competency Mapping (CCM) tool by which the competency requirements of specific or generic jobs are analysed in order to obtain person-job-matching results.

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. I 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. Vary 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.

So, let us return to a rephrasing of question 2 to be answered for the CPP specifically; the one which users of the assessment want answered.

Do the CPP assessment results and subsequent interpretations show the expected relationships with other theoretically relevant scores, behaviours, and outcomes?

The summary of research findings on the CPP provides a body of evidence of meaningful relationships with assessments of intellect, ability, emotional intelligence, and personality. However, while these relationships are of interest in terms of indicating that the CPP attributes are related in various ways with these other psychological attributes, they remain 'an aside' to the primary focus of investigation; that of the proposed consequences of using a CPP as an assessment tool.

Those consequences revolve around the preferences and scores associated with an individual's performance on the CPP which are claimed to result in particular kinds of outcomes, of interest to the purchaser/user of the assessment. To put it colloquially, "*does it do what it says on the box?*".

Some practitioners/psychologists might want to refer to this as assessing 'predictive validity', but that is too-specific terminology. Predictive validity is associated with quantifying the associations between test scores and specific outcomes; an essential component of any consequential analysis for a test (some results reporting routine predictive accuracy of the CPP can be found in the research manual on our website).

From the website for the product, we claim the CPP will be useful for:

Identification of potential
Succession planning
Career guidance
Personal and team development
Selection and placement
Organisational development and capacity building
Anchoring competency assessments
Intellectual Capital Management

However, when we wish to address these kinds of claims made on our website, we are attempting to justify 'knowledge claims' about assessment utility. That is, we are making claims concerning practitioner 'utility', not claims about the precise predictive accuracy of any single score or attribute magnitude. Sometimes the two will coincide; however, it is possible to consider evidence for utility independent of evidence of 'predictive accuracy'.

Remember what I said above:

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 known other information about a testee. 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 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 world-wide 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 a new whitepaper 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, 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 neighboring 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 important for CPP validation strategy 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 CPP validation into new but meaningful territory.


Laboratory experimentation using task designs which isolate specific cognitive processes that can provide equivalency evidence for the processes which the test author has hypothesized as being assessed by the CPP. These will be conventional scientific experimental studies; where the focus is solely on establishing equivalency rather than the more usual moderate-to-low concurrent associations.


We are asking the practitioners and users of our 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 CPP added value to your work and/or your organization?"

We are not interested in detail, simply whether or not users have found the CPP 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.

3 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 vs 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?

To acquire such data, we need to work with those who formed judgements and made decisions using the information supplied by the CPP and reports constructed from it. This also means working with the organizations in which the consequences of those CPP-related decisions have been played out. It means mapping trajectories, outcomes, and events onto CPP and perhaps practitioner-related "indicators".

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 CPP use. 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 the case of the CPP, the above mentioned approaches have to some extent been used in validating the tool and in refining the scoring and reporting algorithms. The approach, however, largely entailed qualitative action research rather than a rigorous approach aimed at creating an evidence base.

During the first 20 years of the CPPs application, a large number of such studies have been conducted. Examples include the assessment and comparison of CPP profiles of people between and within career and cultural groups such as Accountants (in Assurances, Tax and Consulting), Actuaries, Credit managers (Corporate versus Commercial Credit managers in Banking), Programmers (using rule-based versus object oriented approaches), Sales consultants, Call centre staff, Business Consultants, HR practitioners, Managers and Executives across industries and cultures and Engineers in various fields and across cultures, to name but a few.

Large numbers of pilot studies and case studies were also conducted in almost a thousand corporates across industries – often as part of a sales strategy or to ensure continued buy-in of corporate clients. In addition, long term corporate clients also tend to request follow up on succession plans and job performance of CPP selected staff.

Feedback from those being assessed, HR practitioners and business executives is standard practice in the case of the CPP. The developer of the CPP has personally watched hundreds of people doing the CPP, given feedback to thousands of test-takers, and guided in-house corporate studies to establish the value add of the CPP within organisations across industries. Again, these studies were mostly conducted in a qualitative manner, aimed at ensuring value add to the clients (which are the test-takers, the consultants and corporates involved), rather than for the purpose of creating an evidence base for the tool.

The performance of employees who were CPP assessed, selected and placed have thus been evaluated. The results of some of these studies are documented in the CPP Research manual in the section on Predictive validity. One example is a 5 year longitudinal study that was conducted in an Accounting firm with 752 trainee accountant bursary holders. The organisation was particularly interested in the identification of potential in accounting graduates from socio-economically disadvantaged backgrounds. Before using the CPP as a selection tool, the firm showed a 4% pass rate on the CA exam for accounting students from disadvantaged backgrounds. The group of trainee accountants that were selected using the CPP, however, showed a 64% pass rate – well above the national average of 50% (for both students from both disadvantaged and socio-economically advantaged groups). Such findings ensured long term buy-in into selection practices using the CPP. Many such studies have been undertaken and are in progress.

Not only are the performance of CPP selected managers and executives followed up over time, but at a number of organisations, the CPP is used to assess the complete management and executive team to predict high fliers and under performers. The CPP findings are then compared to the everyday functioning of these individuals as assessed via 360 degree techniques, and the perception of the top executive. In certain such studies the CPP results are integrated with that of alternative assessment tools such as the Value Orientations (VO) and the Motivational Profile (MP) and the positions involved are analysed using the Contextualised Competency Mapping (CCM) tool. The results are provided and explained to the top executive to inform their decision regarding the further use of the CPP and other tools and future assessment practices.

The ongoing qualitative action research on the CPP is an important complement to the approaches outlined above; part of the mix of strategies evolved in order to construct a sound evidence base for the validity of the tool.

In Conclusion

- It is hoped that the reader now has a better appreciation of how we, in Cognadev, are approaching matters of reliability and validity for the CPP.

- The routine "recommended practice" analyses that constitute conventional approaches to the estimation of reliability and validity for any psychological assessment has already been completed and reported upon in our research manuals.

- But, for the CPP, we need to acquire evidence that speaks directly to reliability and validity for a performance-based assessment which contraindicates a typical retest strategy and for which validity is a function both of interpreter and objective test scores and designations.

- Given the further complication of the mix of non-quantitative and preference/typologies assessed by the CPP, it is clear that the assessment of validity must be approached more imaginatively and effectively than simply correlating test scores with generic supervisor ratings.

- We want to answer that simple question "*does it do what it says on the box?*" with clear answers and appropriate evidence.

- Aspiring to provide a forest of 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.

Technical Appendix 1: Measurement

1: Quantitative Measurement

Michell (1990), p.63 ...

> "Quite simply, measurement is a procedure for identifying values of quantitative variables through their numerical relationships to other values. Take a simple example. We wish to know the length of a timber beam. This may be done by relating its length to that called a meter. It is to be found r meters long (where r is some real number). Here r is the ratio of the length of the beam to that of a meter and this fact enables the length of the beam to be characterized. More generally, in measurement some (unknown) value of a quantitative variable is identified as being r units. A unit of measurement is simply a particular value of the relevant variable. It is singled out as that value relative to which all others are to be compared. Let the unit be Y and let the value to be measured be X. Then a measurement has the form $X = rY$.... Measurement requires the development of procedures whereby values X and Y may be brought into comparison and their ratio assessed. Such procedures are the methods of measurement"

Michell (2001), p. 212 …

> "Measurement, as a scientific method, is a way of finding out (more or less reliably) what level of an attribute is possessed by the object or objects under investigation. However, because measurement is the assessment of a level of an attribute via its numerical relation (ratio) to another level of the same attribute (the unit selected), and because only quantitative attributes sustain ratios of this sort, measurement applies only to quantitative attributes. Psychometrics concerns the measurement of psychological attributes using the range of procedures collectively known as psychological tests. As a precondition of psychometric measurement, these attributes must be quantitative".

What is immediately apparent is that this definition is absolutely clear, technical, and precise. It introduces the concept of a "**quantitative variable**" (one whose values are defined by a set of ordinal and additive relations). Further, such variables require a **unit of measurement** to be explicitly identified, such that magnitudes of a variable may be expressed relative to that unit. Thus, as stated in the second passage, "**measurement applies only to quantitative variables**". Yes, this is a narrow definition for measurement, but it is unambiguous and technically specified as we shall see below.

2. Quantitatively Structured Variables

A variable is anything relative to which objects may vary. For example, weight is a variable, different objects can have different weights, but each object can only possess one such weight at any point in time. A quantitative variable satisfies certain conditions of ordinal and additive structure. For example, weight is a quantity because weights are ordered according to their magnitude, and each specific weight is constituted additively of other specified weights. Likewise lengths.

Specifically (from Michell, 1990), p. 52-53) …

> "The first fact to note about a quantitative variable is that its values are ordered. For example, lengths are ordered according to their magnitude, 6 meters is greater than 2 meters, and so on. Similarly the values of other quantitative variables are ordered according to their magnitudes.

The nine conditions for measurement

The familiar symbols, "≥" and ">" will be used to denote this relation of magnitude, "≥" meaning "at least as great as", and ">" meaning "greater than". Also the symbol "=" will be used to signify identity of value.

Let X, Y, and Z be any three values of a variable, Q. Then Q is ordinal if and only if:

1) if $X \geq Y$ and $Y \geq Z$ then $X \geq Z$ **(transitivity)**

2) if $X \geq Y$ and $Y \geq X$ then $X = Y$ **(antisymmetry)**

3) either $X \geq Y$ or $Y \geq X$ **(strong connexity)**

A relation possessing these three properties is called a simple order, so Q is ordinal if and only if ≥ is a simple order on its values. All quantitative variables are simply ordered by ≥, but not every ordinal variable is quantitative, for quantity involves more than order. It involves *additivity*.

Additivity is a ternary relation (involving three values), symbolized as "X + Y = Z". Let Q be any ordinal variable such that for any of its values X, Y, and Z:

4) $X + (Y + Z) = (X + Y) + Z$ **(associativity)**

5) $X + Y = Y + X$ **(commutativity)**

6) $X \geq Y$ if and only if $X + Y \geq Y + Z$ **(monotonicity)**

7) if $X > Y$ then there exists a value of Z such that $X = Y + Z$ **(solvability)**

8) $X + Y > X$ **(positivity)**

9) there exists a natural number n such that $nX \geq Y$ $(\text{where } 1X = X \text{ and } (n+1)X = nX + X)$ **(the Archimedian condition)**

In such a case the ternary relation involved is additive and Q is a quantitative variable".

These nine conditions were stated by J.S. Mill in 1843, and later by Hölder (1901) within his exposition of the axioms of quantity. However, as Michell (1999) points out, the influence of Euclid's theory of magnitudes is present throughout the historical development of the physical sciences, and especially within Newton's Principia of 1728. In short, these are the conditions for quantitative measurement which characterise measurement within the natural sciences.

3. Numbers and their status

Up to now, it has been possible to regard the properties of measurement in isolation of the numbers used to represent magnitudes. However, this third issue is fundamental to an understanding of measurement. It is also perhaps the key to understanding measurement in its wider context. A representational theory of measurement in its broadest sense states that measurement requires defining how an empirical relational system may be conjoined with a number system in order to represent magnitudes of empirical quantities using these numbers. An empirical relational system like weight possesses an ordered structure with the relations defined as in section 2 above. For example, if a class of objects that possess the attribute weight can be compared to one another with a relation such as "being at least as heavy as", then the weights standing in this relation to one another are said to constitute a relational system.

In essence, a comparison operation is required to take place between all objects in this system in order to determine whether the relation holds for any two such objects, and to observe whether the properties of the relations expressed in section 2 above can also be observed using the objects that are said to possess weight. A numerical relational system is one in which the entities involved are numbers, and the relations between them are numerical relations. An example of a numerical relation is the set of all positive integers less than say 1000, with the relation of "being at least as great as". Each number can be compared to another and a determination made as to whether the relation holds for that pair. In fact, the same relations as expressed in section 2 can also be applied to such a number system (all positive integers). We can also apply such relations to real numbers, and observe the properties of the same relations but now using continuous quantities rather than discrete values. So, in the case of weight, the numerical representation of weight is achieved by matching numbers to objects so that the order of weights of objects is reflected in the order (magnitude) of the numbers.

The question which now arises is that of the status of numbers. If we treat numbers as an abstract system of symbols, that can be assigned as and how a person decides they should be used to represent objects within an empirical relational system, then we have representationalism in the manner of Stevens (1951) theory, p. 23 …

> "in dealing with the aspects of objects we can invoke empirical operations for determining equality (the basis for classifying things), for rank ordering, and for determining when differences and ratios between the aspects of objects are equal. The conventional series of numerals – the series in which by definition each member has a successor – yields to analogous operations: We can identify members of the series and classify them. We know their order as given by convention. We can determine equal differences, as 7-5=4-2 and equal ratios, as 10/5 = 6/3. This isomorphism between the formal system and the empirical operations performed with material things justifies the use of the formal system as a model to stand for aspects of the empirical world".

Thus, any numerical modelling of an empirical system constitutes measurement. Stevens (1959) stated the more familiar exposition of this statement of measurement as *the assignment of numbers to objects by rule* and that (p. 19) … "provided a consistent rule is followed, some form of measurement is achieved". This seems a reasonable statement on the surface, and has become the *de facto* definition of measurement for psychologists. **But, it is deeply flawed.**

What Stevens did was to remove the status of a numerical relation system consisting of the real numbers as an empirical system in its own right. Up until the 1950s, numbers were considered to constitute an empirical relational system in their own right. The system was self-contained, logical, possessed the required ordering relations that constitute both ordinal and additive operations, and, in the theory of continuous quantity, sustained the necessary ratios necessary for such a theory. In short, both in the manner that scientists used them, as well as in their existence as a relational system, numbers were considered as empirical facts, not abstract entities.

The existence of the empirical relations was presumed logically independent of the numerical assignments made to represent them. In order to assign a numerical system to an empirical relational system, it was required that the empirical relations could first be identified without necessarily assigning numbers to objects within the system. It was a prior requirement that whether or not an empirical relation possesses certain properties was a matter for empirical, scientific investigation. As Michell (1999), p. 168 states …

> "Simply to presume that a consistent rule for assigning numerals to objects represents an empirical relation possessing such properties is not discover that it does; it is the opposite".

For, what Stevens proposed is that it is not the independently existing features of objects *(the properties or relations inherent within objects)* that are represented in measurement, but it is the numerical relations imposed

by an investigator which determine the empirical relations between objects. When stated like this, it is obvious to even the most disbelieving reader that this is not how measurement in the natural sciences has ever functioned – neither is it a rational course of action for constructing and making measurement.

When one considers the real number relational system defined within the continuous theory of measurement to be an empirical fact (Michell, 1994), and that the conjoining of this system to an empirical relational system *(also considered to be a putative or actual fact by an investigator)* is an empirical hypothesis rather than an assertion by an investigator, then the representationalism espoused by Stevens and psychologists since 1951 is seen to be an impediment to any form of scientific investigation, and not as Stevens saw it, a different kind of measurement construction that was uniquely applicable to the social sciences.

It is of interest to note that even Alfred Binet discounted his "intelligence tests" as measurement instruments (Michell, 2012b):

> "**Abstract**: In a comment, hitherto unremarked upon, Alfred Binet, well known for constructing the first intelligence scale, claimed that his scale did not measure intelligence, but only enabled classification with respect to a hierarchy of intellectual qualities. Attempting to understand the reasoning behind this comment leads to an historical excursion, beginning with the ancient mathematician, Euclid and ending with the modern French philosopher, Henri Bergson. As Euclid explained (Heath, 1908), magnitudes constituting a given quantitative attribute are all of the same kind (i.e., homogeneous), but his criterion covered only extensive magnitudes. Duns Scotus (Cross, 1998) included intensive magnitudes by considering differences, which raised the possibility (later considered by Sutherland, 2004) of ordered attributes with heterogeneous differences between degrees ("heterogeneous orders"). Of necessity, such attributes are non-measurable. Subsequently, this became a basis for the "quantity objection" to psychological measurement, as developed first by Tannery (1875a,b) and then by Bergson (1889). It follows that for attributes investigated in science, there are three structural possibilities: (1) classificatory attributes (with heterogeneous differences between categories); (2) heterogeneous orders (with heterogeneous differences between degrees); and (3) quantitative attributes (with thoroughly homogeneous differences between magnitudes). Measurement is possible only with attributes of kind (3) and, as far as we know, psychological attributes are exclusively of kinds (1) or (2). However, contrary to the known facts, psychometricians, for their own special reasons insist that test scores provide measurements. "

Klaas Sijtsma (2012) summarises the dilemmas for psychologists when he states in an article entitled "*Psychological measurement between physics and statistics*", (p.786-787):

> "Psychologists' attempts at measuring psychological attributes, such as cognitive abilities, personality traits, and attitudes, have met with different kinds of criticism. Blinkhorn (1998) mentioned the limited applicability of psychometric results in practical testing, Lumsden (1976) noticed the tendency in psychometrics to focus on old ideas and indulge in small and irrelevant technical innovations, and Michell (1999) discussed psychologists' uncritical acceptance of the assumption that attributes can be quantified and hence measured without bothering to demonstrate the validity of this assumption for specific attributes. Gould (1981/1996) even questioned the possibility of measuring psychological attributes at all.
>
> Theory & Psychology has published two critical discussions on the issue of how measurement in psychology should be done. On the one hand, Michell (2000, 2004, 2008) and Kyngdon (2008a, 2008b) take the position that psychometrics is inadequate for psychological measurement and must be replaced by additive conjoint measurement (ACM; Luce & Tukey, 1964). I call this the physicist perspective and argue that it asks too much of contemporary psychology: its serious implementation would lead to a standstill in psychological research. On the other hand, Borsboom and Mellenbergh (2004) and Borsboom

and Zand Scholten (2008) argue that modern psychometrics, in particular item response theory (IRT; Van der Linden & Hambleton, 1997), already successfully facilitates psychological measurement. This is the statistical perspective, which I will argue makes the mistake of confusing the prescriptive structure of a statistical measurement model with the theoretical structure of the attribute of interest.

My main problem with the two perspectives is that they leave psychology out of the equation. **Specifically, I believe that meaningful measurement is possible only if enough is known about the attribute so as to justify its logical operationalization into prescriptions from which a measurement instrument can be developed.** An immense problem in psychology is that theories about attributes are often not precise enough to justify a logical operationalization. Both the physicist and the statistical perspective have little eye for this problem. The physicist perspective assumes that psychological theories about attributes can reach a high level of precision comparable to that of theories about physical attributes, but for the time being this is an unattainable goal. The statistical perspective assumes that attributes have the structure of IRT models, but ignores that this assumption is not based on well-developed substantive theories about specific attributes and that, except for rare cases, there is no compelling evidence for the assumed congruence. ”

An important overview is provided by Sherry (2011) of how measurement has been constructed over time within physics, and how this history of initial observations and subsequent instrument calibration work might shed light on how psychologists could approach measurement of psychological attributes. His article is entitled “*Thermoscopes, thermometers, and the foundations of measurement*”:

“**Abstract**: Psychologists debate whether mental attributes can be quantified or whether they admit only qualitative comparisons of more and less. Their disagreement is not merely terminological, for it bears upon the permissibility of various statistical techniques. This article contributes to the discussion in two stages. First it explains how temperature, which was originally a qualitative concept, came to occupy its position as an unquestionably quantitative concept. Specifically, it lays out the circumstances in which thermometers, which register quantitative (or cardinal) differences, became distinguishable from thermoscopes, which register merely qualitative (or ordinal) differences. I argue that this distinction became possible thanks to the work of Joseph Black, ca. 1760. Second, the article contends that the model implicit in temperature’s quantitative status offers a better way for thinking about the quantitative status of mental attributes than models from measurement theory”.

For those who wish to enquire further about CPP measurement/validity matters, please contact Paul Barrett directly at paul@cognadev.com.

Technical Appendix 2: The VRAG construction methodology

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) ... see also Rice (1997).

This is an assessment used to predict the risk of violent recidivism in offenders seeking parole, the risks involved in transfer of 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 from case-records on 618 Canadian male patients who had been released from their high security mental health institution during a period prior to and up to a final date of April 1988 *(on average, they were at risk of recidivism for a little under 7 years duration)*. They also tracked the violent offence recidivism of these patients post-discharge - and reported the initial findings in 1993-4. What they were trying to do was isolate those variables *(derived from examining patient record information)* that were key predictors of recidivist behaviour. Their first task was to determine the key predictors. Their second task was to form a scale of these predictors, assign weights to them (reflecting their importance to the prediction), and subsequently develop a cumulative scale of risk (the VRAG).

Initial empirical exploration of the candidate information most predictive, in combination, of violent recidivism over a fixed period produced 12 predictor attributes. Note, all but #12 are a mixture of biodata and clinical diagnoses; predictor #12 is a psychological assessment, a rating checklist whose ratings on affective and behavioural attributes are provided either by trained raters from patient/offender records or by offender interview with a trained forensic clinical psychologist or forensic nurse. These predictor attributes were:

1) Lived with biological parents to age 16 *(except for death of Parents)*
2) Elementary school adjustment
3) History of alcohol problems
4) Marital Status
5) Criminal History Score for Non-Violent Offences
6) Failure on prior conditional release
   (includes parole/probation violation or revocation, failure to comply, bail violation, and any new arrest while on conditional release)
7) Age index offence *(at most recent birthday)*
8) Victim Injury *(for index offense; the most serious is scored)*
9) Any female victim *(for index offense)*
10) Meets DSM III Criteria for any Personality Disorder
11) Meets DSM-III Criteria for Schizophrenia
12) Psychopathy Checklist-Revised score

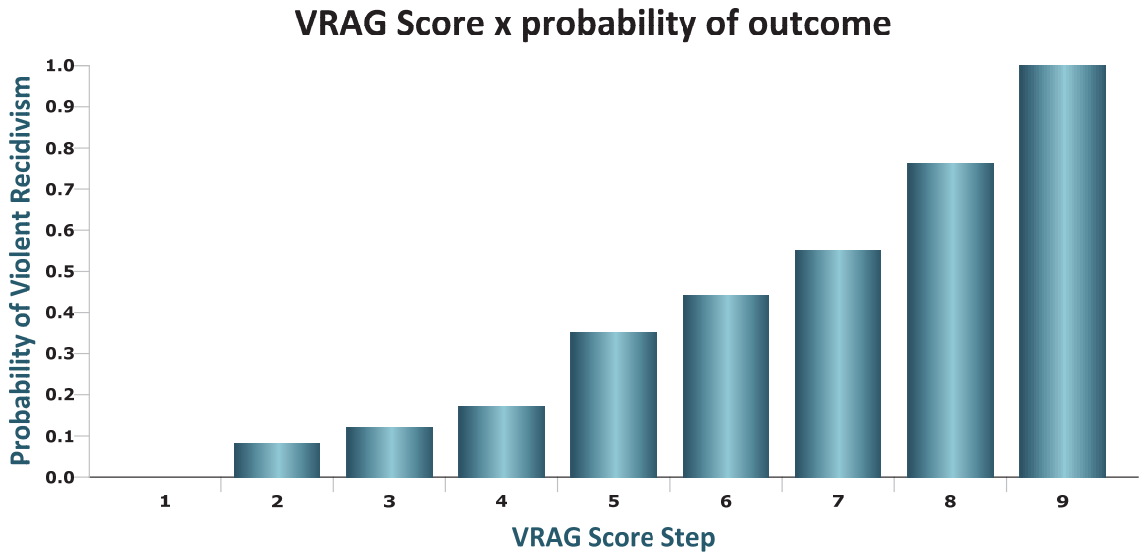
The process of forming the risk assessment scale is taken from the Webster et al (1994) test manual, pp. 33-34:

"Harris et al. (1993) were interested in examining the relative performance of subjects in a fairly wide array of risk levels. The investigators used an adaptation of Nuffield's (1982) method to develop an instrument which would break the sample into subgroups of varying risk levels. Recidivism rates for each score or range of scores on each of the 12 variables were determined. Then, each variable was accorded a weighting of +1 or -1 respectively for every plus or minus 5% difference from the mean recidivism rate of 31%. For example, it was determined that subjects who had married at some point in their lives had a recidivism rate of 21%. With a difference of minus 10% from the mean recidivism rate of 31%, the variable "ever married" was accorded a

weighting of -2. It was also determined that those who had never married had a recidivism rate of 38%. They received a score of +1 since their recidivism rate was over 36% (31 +5). Because it had the strongest correlation with violent recidivism, the highest possible weighting was given to the PCL-R where risk scores from -5 to +12 were possible.

Using all 12 variables, scores ranged from -27 to +35. These scores were then divided into 9 equal-sized steps. Men in Step 1, with very low scores, would be expected to be unlikely candidates for violent failure. The probability would be anticipated to be zero or near zero. Men whose scores fell into Steps 8 or 9 could be expected to fail with reasonable confidence. The probability would be expected to be 1.0 or close to it."

The test scores (1-9) provided the following risk-propensity graph:



VRAG Score	1	2	3	4	5	6	7	8	9
VRAG Prob.	0.0	0.08	0.12	0.17	0.35	0.44	0.55	0.76	1.0

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 I 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.

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.