Strategic Whitepaper

November 2011

Using Psychometric Test Scores:
Some Warnings, Explanations, and Solutions for HR Professionals.

Issue: When outcomes are critical

Using norm-based test scores to convey information about critical potential outcomes is wrong, and legally indefensible unless the norm group is conditioned upon the criterion of interest. Even then, there is little or no utility in this approach. The application domain of employee safety tests is used as an exemplar.

Advanced Projects R&D Ltd.

- Psychological Test Design and Construction
- Predictive Analytics and Profile/Classifier Construction
- Independent Scientific Evaluation/Optimization of Psychological Test Validity

19 Carlton Road, Pukekohe, Auckland 2120, New Zealand

T	+64-9-238-6336
M	+64-21-415625
F	+64-9-280-6121
W	www.pbarrett.net
E	paul@pbarrett.net
Skype	pbar088

When Test Scores are Critical

In some organizational applications of psychological assessment, the test scores for an individual might be considered critical. That is, employment/promotion decisions about a candidate or incumbent may be made directly on the basis of that test score. A couple of areas where this is clearly the case is with integrity/security personnel testing and the assessment of employee safety using psychological attributes as 'indicators of potential risk'.

The outcomes of employee dishonesty, theft, shrinkage, or causing accidents/incidents in the workplace are usually considered critical because the consequences of each can be traumatic, impact on other employees' situations/health, and be financially costly to the organization.

The use of assessments in these domains is generally predicated upon the accuracy of a test to predict an adverse outcome, so that threshold or cut-scores/regions of interest might be used to screen individuals prior to employment, training, or deployment.

Unlike those application areas where psychological assessment information is subjectively combined with other sources and interpreted by one or more members of a selection panel, critical outcome tests produce scores which need no interpretation as their magnitudes are related directly to the probability of occurrence of the adverse outcome. And these 'tests' may actually be composite assessments of many attributes configured into an optimal profile classifier, where the 'score' is not a simple sum of unit-weighted items, but a weighted composite.

What sets this kind of test design, construction, and calibration process apart from standard psychometric tests is that they are computationally optimised for predictive accuracy of the criterion of interest.

Using normative scores with critical outcome tests makes no sense at all, except where a normative group is defined explicitly in terms of the outcome; not in terms of employee group etc. But, even here it makes no real sense to use 'normative' scores. Critical test scores **must** convey likelihood of outcome in order to justify their use.

Employee Safety Assessment

For example, in a safety assessment, should one wish to express a classifier or test score relative to a safe *(no recorded incident)* or unsafe *(at least one recorded self-cause incident)* group, the resulting normative score at least preserves the relativity with respect to the criterion. But, expressing a score relative to an employee group, without first conditioning upon incident-causation confounds the criterion with the group membership e.g. in a supervisor group, some of them may have actually caused incidents and will be included in your norm group.

For HR who want a test score which indicates the likelihood of an individual of causing an incident *(being cognisant of the assumptions in attempting to use a generalization to predict a particular; Faust, 1997),* there is zero value in knowing that a person scores at the xth percentile relative to a group of supervisors who may actually be a mixture of 'incident causers' and 'non-incident causers'. But, a score which relates directly to the probability of that individual being likely to cause an incident over a specified duration *(within 3 months, 6 months, one, two, or three years say)* provides the kind of information upon which decisions about 'risk' might be made more rationally.

Tests constructed for such 'decision-oriented' applications **must** possess an actuarial evidence-base; not a correlational-validity coefficient base. Observed prediction error, and the kinds of errors which might be made (false-positives/false negatives) are critical knowledge for HR to understand in terms of balancing risks.

Two examples of critical assessments which enable decision-makers to utilize test scores as absolute indicators of risk.

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 patient-oriented variables (case-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 an additive 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:

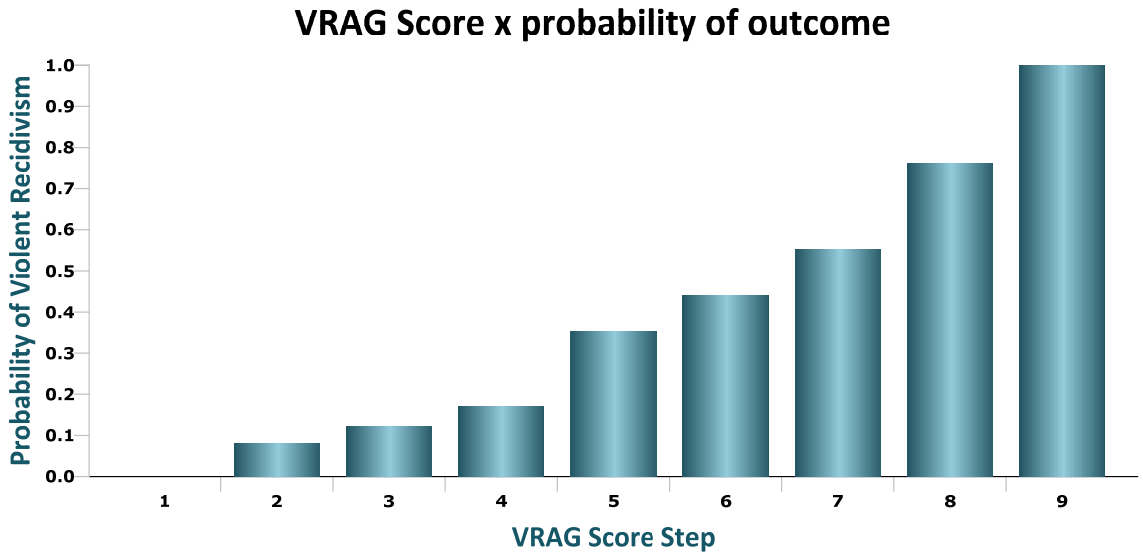
1) Lived with biological parents to age 16 *(except for death of Parents)*
2) Elementary school adjustment
3) History of alcohol problems
4) Marital Status
5) Criminal History Score for Non-Violent Offences
6) Failure on prior conditional release
   (includes parole/probation violation or revocation, 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 probablility 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. Normative scores would impart zero benefit. All the information required to interpret the score is given by the probability of occurrence of the criterion outcome. Subjectivity of interpretation of any new individual’s scores is required when considering two questions:

a) 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.

b) what level of risk constitutes an unacceptable level?

There is no requirement or need for a 'narrative report', some kind of interpretation of 'risk competencies', or trying to interpret the correlation between VRAG scores and recidivist risk (0.45) in such a way that a standard interpretation can be afforded to each VRAG score-level. Note also that scoring is *algorithmic*, carefully designed and calibrated against the criterion of interest.

Given the data at hand, a score of 9 indicates 100% of an adverse event occurring within a 7 year post-release period.

There are other problems associated with using actuarial assessments, as partly indicated in the two questions requiring an answer above *(and as indicated in Faust, 1997),* but the key feature is that a criterion-calibrated evidence-base is so obviously optimal for critical outcome tests.

This outstanding example of test development earned Marnie Rice an APA award in 1997 for a Distinguished Contribution to Research in Public Policy.

The test results were cross-validated many times (e.g. Rice and Harris, 1997; Harris, Rice, and Cormier, 2002; Douglas, Yeomans, and Boer, 2005; Quinsey, Harris, Rice, and Cormier, 2006).

References

Faust, D. (1997). Of science, Meta-Science, and clinical practice: The generalization of a generalization to a particular. *Journal of Personality Assessment*, 68, 2, 331-354.

Harris, G.T., Rice, M., & Cormier, C.A. (2002). Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law and Human Behavior*, 26, 4, 377-394.

Harris, G.T., Rice, M.E., & Quinsey, V.L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior*, 20, 315-335.

Nuffield (1982). Parole Decision-Making in Canada: Research Towards Decision Guidelines. Ottawa, Ontario: Ministry of Supply and Services, Canada.

Quinsey, V.L., Harris, G.T., Rice, M.E., & Cormier, C.A. (2006). Violent Offenders: Appraising and Managing Risk. American Psychological Association.

Rice, M.E. (1997). Violent offender research and implications for the criminal justice system. *American Psychologist*, 52, 4, 414-423.

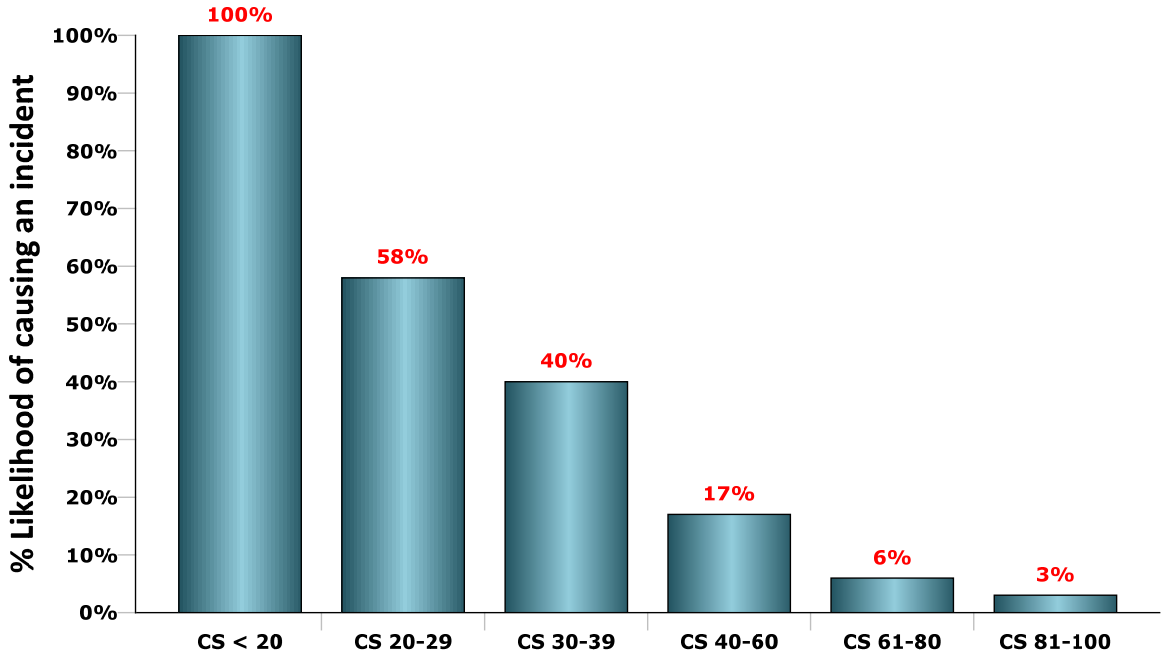
Rice, M.E., & Harris, G.T. (1997). Cross-validation and extension of the Violence Risk Appraisal Guide for child molesters and rapists. *Law and Human Behavior*, 21, 2, 231-241.

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.

An Organizationally-Relevant example: **Employee Safety Assessment**

The consultancy brief was to develop an assessment of employee safety for an international infrastructure company using a mixture of preferences, attitudes, and personality attributes considered relevant to the issue of safety at work. All employees in the organization, from Directors to Operatives were required to be assessed using the same assessment, and assigned a risk-level *(approximately 2000 initially, then another 10,000+ in addition to the use of the assessment for new employee selection).* Test completion was required to be undertaken in less than 10 minutes. The test was required to possess an evidence-base for its predictive accuracy of self-caused incidents. The consultancy team had access to three consecutive years of incident records for the workforce.

The final cross-validated solution *(two samples of employees numbering 1000+ in total)* utilized a small number of attributes, with a classifier score created using a production-rule threshold-scoring algorithm. Median assessment completion time was **5** minutes. The classifier score (**CS**) ranged between 0 and 100. Optimized cut **CS** predictive accuracy was 72%, with balanced accuracy across incident (72%) vs no-incident groups (71%). But, as with the VRAG, what we designed was an assessment which would provide a direct prediction of the likelihood of the critical outcome; in our case the causing of an incident in the workplace. The risk-propensity graph is:



When cross-related against the distribution of **CS** magnitudes, selection ratios and likely frequencies of cases to be expected within each score range can be computed, which helps HR model the likely impact of the test when used for selection or development of 'at risk' employees.

This kind of test-design and local validity evidence-base is powerful, not just because it allows HR to rationally plan workforce strategies using the various rates identified in the calibration analyses, but also avoids the impossibility of trying to interpret 'normed' risk scores which are disconnected from the outcome of interest.

Of course, this kind of test design and solution does require the assessment developer has access to criterion data of sufficient quantities and quality to permit formal local calibration.