Strategic Whitepaper

November, 2011

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

Question: How might you respond when a candidate, employee, or union notifies you of a claim of unfair or negligent practice against you, asserting that the selection procedures utilized resulted in their failure to be employed, retained *post-downsizing*, or promoted? When examined very closely, the usual 'best practice' responses from HR and I/O psychologists might not work.

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

The Scenario

As the HR executive responsible for initiating psychometric assessment of incumbent employees, who are part of a cohort of potential candidates for fast-tracking to leadership development positions within the corporate, you have set up a comprehensive assessment of leadership-potential encompassing cognitive ability, competencies, personality attributes, and dark-side derailer attributes.

The attributes chosen as reflecting the characteristics required for future leaders in your company were a result of a comprehensive job analysis by yourselves, in conjunction with assistance provided by a consultant and the R&D department of an established professional test publisher.

When you purchased the assessments from the test publisher, you asked for confirmation that the assessments are *reliable*, and *valid*. You needed assurance that the assessments 'measure' what the test publisher claims they measure, and that what is measured is predictive of "Leadership Propensity".

The test publisher most likely made available test manuals in which one or more studies, conducted by them or published by academic researchers, were reported. Alternatively (or in addition) you would be shown annual validity 'PR-briefing' documents such as those from SHL (http://www.shl.com/assets/SHL-BOS2011.pdf) or Hogan Assessments ... (http://www.hoganassessments.com/sites/default/files/brochures/Hogan_ROI.pdf). Then again, the publisher sales consultant might simply have assured you verbally that they have all this information "in house/taken care of" should you ever wish to see it.

You deploy the assessments, interview candidates, review performance records and discuss matters with relevant supervisors. You then arrive at a decision which results in selecting a cohort of 20 candidates from the 100 assessed. It was an efficient, well-documented, and well-run process.

However, in the weeks following, several of the rejected candidates begin to claim that the assessment process was unfair, in that although their performance capability had never been questioned in the past by their supervisors, their aspiring career trajectories were now being harmed by what appeared to be a reliance more on the 'psychometrics' than their job performance ratings.

Not satisfied with your efforts to explain how the decision was made, and how the psychometric test scores were used in an overall decision process, they lodge a formal grievance against you (the company) in an appropriate employment court. For them, they see they have nothing to lose as their careers are finished in the company unless they can argue that the selection process was unfair or at least 'potentially negligent and/or inaccurate'. If they can do that they will also look like 'courageous' people who took a stand on unfair practices and won. A nice look for potential ethical and principled leaders. For you, your own career trajectory might be at stake. To lose would mark you down with those who matter as a potentially 'unsound' executive, a bit 'risky'. Nobody will tell you this face-to-face, but you know what is really at stake here; your reputation.

And so it begins.

The 1st line of Defence

The disaffected group assume that we made our decision using the test scores alone, or were disproportionately biased in the importance we gave to them. As our test publishers advise, as well as the relevant international professional societies, we used the psychometric scores only as a component (among many) to help us come to a decision.

This is the standard position promoted by I/O psychologists and test publishers alike, in an attempt to ensure that psychometric test scores are not treated as 'deal-breakers', and thus subject to standards of evaluation beyond the 'merely superficial'. By embedding scores within a complex mix of other information about a candidate, it is assumed that this strategy defuses any claim that the scores might have unfairly influenced the decision-makers.

The arguments deployed here are twofold:

1. Psychometric test scores are like any other kind of information about a candidate. They describe features and attributes of a candidate in ways not described by their job performance, references, supervisor ratings and discussions, and interview-related information. All of this information is carefully discussed and taken into consideration by the selection panel. Sometimes, test scores might well be overridden by cogent arguments from the panel.

The key argument here is that the test scores are like any other kind of information used by a decision-maker. All such information contains error to some extent, and it is the reasoned judgment of the decision-makers which is applied to the mix of information in order to arrive a final decision.

The problem with this line of argument is that it only succeeds in moving the evidence-requirement from the test scores to the behaviour of the decision-makers themselves. In the manner in which decisions about psychiatric patients made using subjective 'clinical judgement' by 'expert practitioners' was eventually rendered inadmissible in many international courts *(largely thanks to the trilogy of books from Ziskin (1981, 1995, Faust, 2011), which provided the means to cross-examine the claimed expertise and decision-making of psychiatric and psychology professionals in court)*, the same approach can be taken with HR decision-makers who rely entirely on their 'expertise and experience' when arriving at selection judgments involving the integration of several information sources. Unless an evidence-base exists for the fairness and accuracy of those judgments, or, the processes involved in the blending of scores, other information, and the chains of logic/deduction by which judgments were made can be made explicit to any interested 3rd party, the HR executive/selection panel is exposed to formal challenge.

Within the clinical domain, *structured professional judgement* (SPJ) is the term used to denote the formal methodology for blending 'expert judgment' with factual information sources (Webster, Douglas, Eaves, & Hart, 1997; Webster, Müller-Isberner, & Fransson, 2002). Clinicians are provided with a formal (structured) rating scheme which incorporates known attributes related to certain outcomes which require a judgment to be made by them, as well as incorporating factual (actuarial) information into the overall judgment. The SPJ decisions are validated in terms of their accuracy against known outcomes. While this specific approach to validation is not completely relevant to HR selection practices, it is nevertheless achievable. To not attempt to do so might be argued by a skilful lawyer as professional negligence or even incompetence.

2 Test scores, especially from personality, motivation, values assessments etc. can be somewhat inaccurate predictors of many relevant specific workplace outcomes. It is unreasonable and unethical to use the scores alone as the means to arrive at selection decisions.

This is an awkward defence. Test scores are often used for pre-screening candidates prior to short-listing. Successful defence of the selection practices cannot rely upon this argument as the paradox of their use in many 'cut-score' screening or profiling applications shows that on occasion their accuracy is accepted as 'veridical'.

However, there is one distinguishing feature of applications which use psychometric test scores for screening purposes. A special kind of evidence-base is usually constructed which provides a form of 'dose-response' relationship between test score/profile score magnitude and outcome prevalence; the user of a test score or score-composite is NOT left having to subjectively interpret its magnitude in the context of a validity coefficient expressed as a 0.3 relationship between transformed test scores and transformed outcome data. *Recall that actual test scores and actual outcomes are not used in a standard validity coefficient correlation coefficient. Variables are 'standardized' in such a way that only the monotonic relationship between magnitudes is retained.*

Explanatory Box #1: A problem with Correlation as Validity

For example, in Box #1, some Customer Focus scores and Job Performance Ratings are provided in the upper table.

Customer Focus scores range from 0 to 36, high scores indicate high customer focus. *Performance ratings* range from 1 to 5, where
1 = unacceptable
3 = average
5 = excellent

The Pearson correlation *(validity coefficient)* between the test scores of *Customer Focus* and *Performance "on the job" Rating* is **0.73**. Larger test scores are clearly associated with higher performance ratings.

The lower data matrix shows a similar correlation (**0.68**) between test scores and performance, but now all but one of the performance ratings are less than average.

	APR&D Ltd. {PB} - Strategic Whitepaper #4: Correlation			
	1 Customer Focus	2 Actual Performance Rating	3 Standardized Customer Focus	4 Standardized Actual Performance Rating
1	12	3	-1.275	-0.682
2	14	2	-1.079	-1.774
3	15	4	-0.981	0.409
4	23	3	-0.196	-0.682
5	32	4	0.686	0.409
6	33	4	0.784	0.409
7	36	5	1.079	1.501
8	35	4	0.981	0.409

The raw scores/ratings are given in columns 1 & 2.
What's actually analyzed is in columns 3 and 4.

	APR&D Ltd. {PB} - Strategic Whitepaper #4: Correlation			
	1 Customer Focus	2 Actual Performance Rating	3 Standardized Customer Focus	4 Standardized Actual Performance Rating
1	12	2	-1.275	0.195
2	14	1	-1.079	-1.365
3	15	1	-0.981	-1.365
4	23	2	-0.196	0.195
5	32	2	0.686	0.195
6	33	2	0.784	0.195
7	36	2	1.079	0.195
8	35	3	0.981	1.755

Only one person was rated "average",
all others were adjudged as *unacceptable*, or *poor* performance

I do not wish to imply that all correlational data will yield such misleading results, merely that unless due care and attention is paid to 'validity' correlation coefficients provided to you, it is possible that detailed analysis of the raw data underlying a validity coefficient might expose just such an issue, with the obvious ramifications in a grievance setting. This assumes, of course, that you can ever be shown the raw data or sufficient diagnostic summary data to ensure any validity coefficient is not spurious. **Barrett View #2** (http://www.pbarrett.net/tbv/Index.html#BV2) outlines a very recent NZ employment-court judgment which now makes this possible. *E.g. someone like myself acting on behalf of a plaintiff can now be provided with the original 'validation' data for expert re-evaluation/re-analysis.*

> Strategic Whitepaper #5 (http://www.pbarrett.net/#whitepapers), entitled "**Using Psychometric test scores: When outcomes are critical**" provides some examples of functional-relationship *(dose-response)* scoring of a test which allows clear interpretation of score magnitude and outcome prevalence. An example from forensic psychology (risk of violent recidivism) and from organizational psychology (employee safety assessment) show what is required to provide test scores to users which possess the required evidence-base to support their credible use in selection scenarios.

Ultimately, deploying ❷ as a 'contributory' argument in the 1st line of Defence is an admission that the selection panel recognize that the test scores are of unknown accuracy in terms of their predictive relation to the outcomes of interest (those associated with successful/desirable Leadership behaviours in the specific organization using such tests). Which in turn begs the question from an employer advocate "Why did you use these tests/ assessments if you now admit you were/are unsure about their accuracy in terms of predicting the kinds of outcomes for which you are making a selection decision?" Stating under adversarial challenge that the test publisher or vendor (consultancy) assured you of the accuracy of the assessments as measures/predictors of those attributes/outcomes relevant to your application, and you proceeded with the selection strategy based upon the veracity and content of those claims, attention shifts to the test vendor representative. They must now attempt to justify the accuracy of the validity information provided to you in court (via affidavit or in-person).

However, this will be problematic if you invoked ❷ as a 'contributory' argument. It is a 'no-win' situation.

Why? Because:

- if the test vendor is *successful* in showing the court that the test scores did indeed possess the kind of validity which would stand expert scrutiny, then they are begging the question of why you ignored that evidence in your selection procedure, and argued that *"test scores, especially from personality, motivation, values assessments etc. can be somewhat inaccurate predictors of many relevant specific workplace outcomes. It is unreasonable and unethical to use the scores alone as the means to arrive at selection decisions."*

- if the test vendor is *unsuccessful* in showing the court their test scores possessed the kind of accuracies which could have been used confidently in the selection procedures, then it leaves you exposed as having used test scores which the vendor has been unable, under cross-examination, to justify as 'fit for the explicit purpose employed'.

Given the reasoning and potential outcome analyses above, the I/O consultant or test publisher/vendor is likely to advise a shift of strategy, employing a completely different approach to defending the selection practice.

The 2nd line of Defence

Modern organizational psychological science has shown that validity generalization via meta-analysis studies and/or synthetic validity coefficients generated by a test publisher render **local validity** *(determining whether the assessment works in your organization)* as obsolete.

This line of defence is a complex mix of sophisticated statistical and logical arguments employed to show that the tests used possess validity, with an evidence-base drawn from meta-analytic sources and/or synthetic validity studies. Test publisher/vendor R&D experts and/or expert academic quantitative psychological expertise will be deployed via affidavit or in-person. This is normally a clever strategy, as those initiating the grievance are unlikely to be able to respond at all to this kind of evidence, without hiring expensive expert assistance of their own. And finding this 'expert assistance' is not a simple matter as those most capable of questioning the defence experts are likely to agree with them because so many 'experts' in psychological assessment defer to credentialed persons rather than credentialed knowledge (Meehl, 1997), and/or because they possess the kind of non-independent relationship with test-publishers or professional organizations which would not look kindly upon their questioning the status quo or received wisdom associated with this line of argument.

▶ Validity Generalization using Meta-Analytically derived validity coefficients. The evidence-base for the accuracy of a test to predict relevant outcomes is provided by aggregate empirical summaries of many individual studies, yielding the best estimates of relevant validity parameters. Because these estimates represent averages computed across many samples from many difference sources, they are considered generalizable across situations and jobs, so that the need for any organization to conduct its own validity study is now rendered obsolete.

Meta-Analysis is the procedure used to calculate the likely 'population' magnitude of a statistical parameter or effect size (e.g. a mean, a correlation coefficient) by averaging many results from samples in which the parameter of interest was originally estimated. Because of sampling error within any single study, the estimated parameter value may not be very close at all to the 'true' value *(being dependent upon sample size, restriction of range of the relevant attributes in any sample, measurement unreliability of the assessments, and representativeness of the sample members to the proposed population).* So, it is possible for many studies examining the same hypothesis to yield parameter values which vary from study to study, making it look like the parameter is in fact an unreliable, non-replicable effect. However, such variation is entirely expected from statistical sampling theory. If the sample estimates are averaged, it is possible to produce a more robust estimate of the 'true' expected population value for a parameter. The famous article by Schmidt and Hunter (1998) provided meta-analytic estimates of validity coefficients (the criteria being *job performance* and *training outcome*) for many methods for selection (personality, ability, interests, values, work-samples, assessment centres etc.); these estimates are now presented as 'generalizable' validity coefficients to clients. That is, the estimates are used to indicate the relationship magnitude any client might expect to find between assessments of particular psychological attributes and job-performance.

Prior to meta-analysis, **local validity** was considered the most appropriate methodology to demonstrate validity for any assessment procedure, as situational specificity was considered the overriding factor which caused validity coefficients to vary from sample to sample. An organization would expect to conduct its own study of whatever it was using in order to establish the evidence-base for the validity of the procedures. Meta-analysis proponents now regard local validity as redundant, and indeed, a very poor substitute for meta-analytic estimates.

Before we respond to this line of argument, it is perhaps worth mentioning the other approach said to render local validity obsolete; **synthetic validity**. This is a method for estimating the validity of an assessment procedure by inferring the task components for a job role via logical analysis and/or job-component-process questionnaires completed by job incumbents. Then, the validity of particular assessments of the attributes identified as constituents of the job components is established. Once these have been acquired, ideally any job role can now be described in terms of a constituent job component, each of which possesses assessments of known empirical validity. In principle, once a job task-analysis has been undertaken and job components identified within a new client organization, a test-publisher/consultancy with access to the job-component validity bank can recommend the best possible assessments for the job-role. In essence, the validity for a particular client job-role is a synthetic composite from a job-component bank (see Johnson, Steel, Scherbaum, Hoffman, Jeanneret, & Foster (2010) for a detailed overview an discussion of this methodology). At this moment in time, there are no established synthetic validity component banks in common commercial use, although the Position Analysis Questionnaire (McCormick, Jeanneret, and Mecham, 1972) component database which fed into the US Department of Labor O*NET job classification database might be considered a potential 'building block' candidate. However, from my own work *(predictive job resolution/discrimination computational experiments with O*NET)*, and from others commenting on the Johnson et al (2010) target paper (e.g. Schmidt and Oh, 2010), an impractical number of discrete attribute assessments are required to accurately predict/discriminate a job-role *(even as a top-five candidate in a ranking scheme of possible predicted jobs)*.

Companies like Hogan Assessments will even produce custom meta-analyses of relevant criterion validity studies on demand, drawn from their own considerable validity study bank. The claim is always that the meta-analytic evidence trumps local-validity or single-study findings, and that the results from the meta-analysis will be generalizable to other organizations/clients wishing to use the tests/attributes evaluated within the meta-analysis. Let us now focus on the validity-generalization line of defence.

There are four avenues of adversarial probing against this kind of defence:

1. The really critical assumption concerns the equivalence of attributes evaluated within a meta-analysis. I.e. the predictors are exactly the same in each constituent sample entering the analysis. Not the same name, but the identical predictors. Likewise, the criteria used as measured outcomes. For example, job performance may be assessed quite differently from company to company. The consequences following failure to meet the assumption is now pretty clear with regard to personality and sales assessments, where constituent samples in a meta-analysis use predictors and/or criteria which are assumed to be the same (*same-name*), but which are actually composed of different test-publisher test-scales or different kinds of 'performance' criteria.

If non-identical predictors are used, or non-identical criteria, then the meta-analysis is averaging study effects which may not only be a function of sampling error, but also of meaning-differences between predictors or criteria. The meta-analysis evidence is now compromised unless evidence can be brought forward showing that while the mix of attribute equivalences may not be perfect, it is sufficiently high to indicate *'good enough for all practical purposes'* equivalence. That means studies showing the virtual equivalence of all the predictor attribute measures to one another; likewise all the 'job-performance' outcome measures if job performance was an outcome variable. This is virtually impossible to provide for most if not all 'performance-based' criterion measures, as the situational specificity alone renders any such study intractable. A selection of some of the evidence indicating the magnitude of this problem is provided below:

1:1 A whitepaper from Barrett and Rolland (2009) detailed the results from various meta-analytic and huge sample (N > 156,000) estimates of published population correlations between two specific Big Five personality test scales, Emotional Stability and Conscientiousness. These are the two most important broad personality factors associated meta-analytically with job performance. Nine sources of published evidence were examined in some detail.

Table 4 from this whitepaper shows the various 'population estimates' reported in each dataset:

Study	Sample Size	Estimated Population Correlation
Ones (1993)	?	0.26
Mount et al (2005)	4,000	0.52
Steel et al (2008)	17, 464	0.33
Meriac et al (2008)	1,009	0.24
Hogan et al (2007) – T1	5,266	0.67
Hogan et al (2007) – T2	5,266	0.70
Hogan HPI Normative US dataset	156,614	0.53
Judge et al (2004)	1,885	0.49
Revelle et al	> 50,000	0.17

The 'population' estimates for the relationship between Emotional Stability and Conscientiousness range from **0.17** through to **0.70**.

1:2 Pace, V.L., Brannick, M.T. (2010) How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*, 49, 7, 669-676.

Abstract

An underlying assumption of meta-analysis is that effect sizes are based on commensurate measures. If measures across studies do not have the same empirical meaning, then our theoretical understanding of relations among variables will be clouded. Two indicators of scale commensurability were examined for personality measures: (1) correlations among different scales with similar labels (e.g., different measures of extraversion) and (2) score reliability for different scales with similar labels. First, meta-analyses of correlations between many commonly-used scales were computed, both including and excluding scales classified as non-Five-Factor Model measures. Second, subgroup meta-analyses of reliability were examined, with specific personality scale as moderator. Results reveal that assumptions of commensurability among personality measures may not be entirely met. Whereas meta-analyzed reliability coefficients did not differ greatly, scales of the ''same" construct were only moderately correlated in many cases. Some improvement to this meta-analytic correlation occurred when measures were limited to those based on the Five-Factor Model. Questions remain about the similarity of personality construct conceptualization and operationalization.

1:3 Woods, S.A. (2009). The comparative validities of six personality inventories. *Proceedings of the Division of Occupational Psychology, British Psychological Society, Annual Conference 2009* – reported in detail in Woods, S.A. (2009). The Structures and Validities of Five Work-related Personality Inventories. *Unpublished Working Paper (Jan 30th, 2009*). The working paper may be requested from S.A. Woods, Aston Business School http://www1.aston.ac.uk/aston-business-school/staff/academic/wop/dr-stephen-woods/).

Abstract

This study examined the structures, convergent validities, and criterion validities of five work-related personality inventories (the Hogan Personality Inventory, the Occupational Personality Questionnaire, the Sixteen Personality Factor Questionnaire, the Personality and Preferences Inventory, Profile Match). A sample of 371 individuals from the UK working population completed various combinations of the five inventories, plus a measure of the lexical Big Five and several criterion scales. Overall, the results indicated sensible and interpretable factor structures for the inventories (with the exception of the Personality and Preferences Inventory), theoretically meaningful convergent validities, and similar criterion validity magnitudes and patterns. It was concluded that these data give confidence in the use of these inventories for research and practice, and reveal important similarities and consistencies in their validities.

A sub-selection of Table 23 on p. 30 presents a summary of the mean correlations for each particular Big Five scale. The table shows the mean correlation between the 5 different tests' measures of each of the Big Five attributes. So, given 5 personality tests, each measuring Extraversion, Woods correlates each test's scores on Extraversion with those from the other tests (pairwise); producing ((5x5)-5)/2 possible correlations. The average is then taken of these correlations.

Table 23. Mean convergence

Domain	All Scales
Extraversion	.38
Agreeableness	.29
Conscientiousness	.31
Emotional Stability	.42
Openness	.29

What these data indicate is that there is only a broad level of agreement between five major work-related personality tests when their scales are re-expressed (where appropriate) as Big Five personality constructs.

Whereas the evidence-bases supporting each questionnaire might well be substantive and excellent, any meta-analysis which incorporated these tests as part of a study looking at the correlation between personality constructs, or even the predictive accuracy of a specified outcome, is going to be influenced by the lack of agreement between test scales.

1:4 Rich, G.A., Bommer, W.H., MacKenzie, S.B., Podsakoff, P.M., & Johnson, J.L. (1999) Apples and apples or apples and oranges? A Meta-Analysis of objective and subjective measures of salesperson performance. *Journal of Personal Selling & Sales Management*, XIX, 4, 41-52.

Abstract
The goal of this study was to examine the relationship between objective and subjective measures of salesperson performance. The results of a meta-analysis of 21 studies with a total sample size of 4,092 participants indicated an overall mean corrected correlation of **.447**, indicating that the two measures shared only about 20% of variance. Although a moderator subgroup analysis found that the corrected mean correlation was somewhat higher in certain situations, the findings generally suggest that objective and subjective measures of salesperson performance are not interchangeable, and that the choice of the most appropriate measure may require a trade-off between accurately tapping the domain of the performance construct and minimizing measurement error.

This is a clear indication that meta-analyses which utilize studies composed of a "same-name" criterion (here "Sales Performance") are likely to yield quite different "population estimates") if the criterion is not exactly the same.

From pp. 41-42 of this paper:
"In the sales management literature, performance has been measured in a variety of different ways. For example, roughly half of the studies (53.3% of the reported correlations) included in Churchill, Ford, Hartley, and Walker's (1985) meta-analysis of the determinants of salesperson performance measured performance using subjective evaluations obtained from managers, peers, or self-reports. The other half (46.7% of the reported correlations) measured volume, sales commissions, or percent of quota. However, in most instances, no attempt was made to explain why one type of measure was used as opposed to another. The implicit assumption appears to be that objective and subjective measures of performance are interchangeable, and that the domain of sales performance is adequately captured by either subjective ratings or objective results. Thus, practitioners and researchers alike are assuming that these two classes of measures are "apples and apples" and that they may be treated interchangeably. "

The same argument might easily be made for that global criterion, "job performance".

Bottom Line:

The constituent samples in meta-analyses put forward as the evidence-base for a client's usage of a test may not survive detailed scrutiny. More specifically, the claim is not that the meta-analysis effects are zero, but that the magnitudes originally provided to the client in support of a sale may be problematic to the extent that they are not trustworthy.

2 Meta-Analytic Estimates are actually not as 'accurate' or veridical as proponents would suggest.

The evidence is drawn from various sources:

2:1 In the response to peer commentary article published by Steel, P., Johnson, J.W., Jeanneret, P.R., Scherbaum, C.A., Hoffman, C.C., Foster, J. (2010) At sea with synthetic validity. *Industrial and Organizational Psychology*, 3, 3, 371-383, the authors indicate on page 376 ...
"Validity generalization avoids this problem if the credibility intervals around estimates are small. Ideally, all the variability would be accounted for, the mean would not vary much across situations, and there would be no need for synthetic validity. Unfortunately, this is not the case, not even for GMA validity coefficients. Consider Hunter and Hunter (1984), who found a **90%** credibility interval ranging from **.29 to .61.** For Hülsheger, Maier, and Stumpp (2007), the credibility interval was from **.30 to .77**, and for Bertua, Anderson, and Salgado (2005), it was from **.09 to .87**. For Conscientiousness, its **90%** credibility interval can be calculated as **.06 to.42**, with even larger ranges for its facets (Dudley, Orvis, Lebiecki, & Cortina, 2006)".

Explanatory Box #2: Credibility Intervals for Meta-Analytic Effect Sizes

From: Millsap, R. (1998) Tolerance Intervals: Alternatives to Credibility Intervals in Validity Generalization Research. *Applied Psychological Measurement*, 12, 1, 27-32.
"A validity generalization study provides estimates of the mean and variance of *true* test validities using the results of many individual validation studies. Schmidt and Hunter (1977) proposed that the estimated mean and variance of the true validity distribution be used to construct an interval, centered at the mean true validity, that contains a specified percentage of the distribution of true validities. This interval was denoted a 'credibility interval', borrowing the concept from Bayesian statistics (Novick & Jackson, 1974)"

Within a sample, the '*true*' validity is that estimated taking into account measurement errors due to unreliability and restriction of range.

The credibility interval is the range within which we might expect to find a validity coefficient with % certainty given the particular samples of data from which the mean and variance of the estimates of *true* effect sizes have been computed.

So, now look at those credibility intervals again for GMA – and consider again the claims you might wish to make about how well "ability predicts performance'.

2:2 The whitepaper referenced above from Barrett and Rolland (2009). The 'population' estimates from meta analyses and huge sample-data for the relationship between Emotional Stability and Conscientiousness range from **0.17** through to **0.70**.

2:3 A target article, peer commentary, and response indicate that within the medical sciences, the assumed accuracy of meta-analysis to yield 'population estimates' has not always been met with success. These articles follow that of LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997) Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *The New England Journal of Medicine*, 337, 8, 536-542.

Hennekens, C.H., & DeMets, D. (2009) The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. *Journal of the American Medical Association,* 302, 21, 2361-2362.

Hennekens, C.H., DeMets, D., Bolland, M.J., Grey, A., Read, I., Vosk, A. &, Sacristan, J.A. (2010) Commentaries and reply to the Hennekens and DeMets Commentary on Meta-Analysis. *Journal of the American Medical Association*, 303, 13, 1253-1255.

In addition, the article by Levine, T.R., Asada, K.J., & Carpenter, C. (2009) Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against non-significant findings. *Communication Monographs*, 76, 3, 286-302.

Abstract

Meta-analysis involves cumulating effects across studies in order to qualitatively summarize existing literatures. A recent finding suggests that the effect sizes reported in meta-analyses may be negatively correlated with study sample sizes. This prediction was tested with a sample of 51 published meta-analyses summarizing the results of 3,602 individual studies. The correlation between effect size and sample size was negative in almost 80 percent of the meta-analyses examined, and the negative correlation was not limited to a particular type of research or substantive area. This result most likely stems from a bias against publishing findings that are not statistically significant. The primary implication is that meta-analyses may systematically overestimate population effect sizes. It is recommended that researchers routinely examine the n-r scatter plot and correlation, or some other indication of publication bias and report this information in meta-analyses.

2:4 There are also some examples of meta analyses yielding untrustworthy results reported in a very recent ScienceNews feature article (Web Edition, Nov 4th, 2011: Odds Are, It's Wrong: Science fails to face the shortcomings of statistics).
http://www.sciencenews.org/index/feature/activity/view/id/335872/title/Odds_Are%2C_Its_Wrong

Bottom Line:

Meta-analysis estimates are not always accurate, and in some instances can yield misleading results. For an individual to rely unquestioningly upon such evidence in support of test or scale usage suggests that the person is not completely up-to-date with the research literature from those areas in which most use is made of this statistical methodology. This is a weakness which can be exploited by counsel.

3 There is an Achilles Heel of meta-analyses which seems to evade those who undertake meta-analyses of attributes predictive of job performance. That is, the relationship between an individual's psychology and eventual job performance will be heavily affected by their particular supervisory situation and other situational features which enter into any validity study.

For proponents of meta-analytic evidence, there seems to be a systemic blindness to what is metaphorically staring them in the face in their trawl through sample libraries and results. On many occasions, studies enter the meta-analysis which perhaps show no relationship or even opposite in sign to the expected/hypothesised relationship. Yes, this is exactly what one might expect in small samples, restricted in range, and with perhaps less than desirable reliabilities of the predictors (or outcome assessments). From a purely statistical perspective, this is not a problem at all. The variation in study effects is reflected in the estimation of variation around a mean effect parameter value.

But, for a single client wanting to know the likely magnitude of an effect if and when they use an assessment, the knowledge that a sizeable proportion of samples achieved the opposite or no-effect when using the assessment has a special meaning all of its own.

It is pointless relying upon a mean effect size, or a credibility interval that ranges between 0.05 and 0.6 as a definitive "evidence-base". These tell a client nothing about the odds of them not seeing a benefit when using the suggested assessment; only that the assessment may have an undetectable effect through to an important one, **if** the meta-analytic estimates were accurate in the first instance.

The reality is, other situational effects can and do negate/augment the expected relationship between predictors and outcomes, not least the effect of an incompetent or superb immediate supervisor. In a perfect world *(which is what error-adjusted meta analytic estimates rely upon)*, all measurement is assumed to be *quantitative (as though we were measuring length, mass, electrical voltage, time etc.)* there is no measurement error, no restriction of range, and all attributes are assumed to be normally distributed. The reality is our 'measurement' is mostly confined to ordered-classes, is non-quantitative, comes with measurement error as standard, restricted ranges due to realistic candidate self-selection for job roles, and most attribute score distributions are skewed (especially applicant data).

A meta-analysis is after all, an aggregate statistical analysis. That's fine when we wish to make statements about what happens 'on average'. But, for a client wishing to justify their use of a procedure based upon what happens "in general" may not be sufficient when tested under adversarial cross-examination. Especially when the means to establish more precisely the predictive utility and validity of the assessment procedure within the client company is available to that client.

Bottom Line:

Meta-analysis estimates are not necessarily accurate indicators of what benefits might accrue to a client with the use of a particular set of assessments. The credibility intervals and proportion of studies not yielding the effect, or indeed an opposite effect also need to be evaluated very carefully. It may turn out that the mean effect size is providing a false impression of the actual likelihood of a company actually seeing an effect of that size.

4 Local validity seems to be what the US courts demand, irrespective of the arguments by proponents of validity generalization.

An important review of cases and judgments is provided in Biddle, D.A., & Nooren, P.M. (2006) Validity generalization vs Title VII: Can employers successfully defend tests without conducting local validation studies? *Labor Law Journal*, 57, 4, 216-237.

Abstract (1st – 3rd paragraphs, p. 216-217)

"The 1991 Civil Rights Act requires employers to justify tests with disparate impact by demonstrating they are sufficiently "job related for the position in question and consistent with business necessity." This requirement is most often addressed by conducting validation studies to establish a clear connection between the abilities measured by the test and the requirements of the job in question. Building a validation defense strategy in such situations requires employers to address the federal Uniform Guidelines on Employee Selection Procedures (1978), professional standards, and relevant court precedents.

In recent years, some employers have attempted to "borrow" validation evidence obtained by other employers for similar positions rather than conduct their own local validation study. This strategy relies on a methodology known as "validity generalization" (VG). Despite the increase in popularity among test publishers and HR/hiring staff at corporations, relying entirely on VG to defend against Title VII disparate impact suits will likely lead to disappointing outcomes because the courts have generally required employers demonstrate local and specific validation evidence where there is local and specific evidence of disparate impact.

The goal of this article is to review Title VII requirements for establishing validity evidence, overview federal and professional requirements for validation strategies (specifically VG), outline how some courts have responded to VG strategies, and conclude by providing recommendations for validating tests that come under Title VII scrutiny".

They conclude (p. 236)

"When choosing between relying on VG evidence to import validity of generic ability tests or conducting local validation studies and/or developing job- and employer-specific tests based on researched job requirements (job analyses, test plans, etc.), the latter option enjoys several major benefits. First, using customized tests is more likely to result in higher validity. In fact, one major study (including 363,528 persons and 502 validation studies) compared the validity differences between written tests based upon job specificity. The results showed that tests highly specific to job requirements demonstrated much higher validity (about double that of "generic" tests), and the results were consistent with both on-the-job and training performance. Another benefit is that custom tests provide a stronger defense if the employer is challenged. Judges and juries (who are almost always novices in testing and statistics) prefer to see, touch, taste, and feel how the job is rationally and empirically related to the test. As pointed out above, only local validation studies can provide local and specific evidence regarding the statistical and practical significance of the test, the type and relevance of the job criteria, and evidence to support the specific use of the testing practice. When employers elect to rely solely on VG studies, they cannot really know that the test is valid for their job or setting".

Interestingly, the most recent issue of the SIOP journal "Industrial and Organizational Psychology" includes a target article: McDaniel, M.A., Kepes, S., & Banks, G.C. (2011) The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, 4, 4, 494-514 with peer commentary. The abstract to this article is:

"The primary federal regulation concerning employment testing has not been revised in over 3 decades. The regulation is substantially inconsistent with scientific knowledge and professional guidelines and practice. We summarize these inconsistencies and outline the problems faced by U.S. employers in complying with the regulations. We describe challenges associated with changing federal regulations and invite commentary as to how such changes can be implemented. We conclude that professional organizations, such as the Society for Industrial and Organizational Psychology (SIOP), should be much more active in promoting science-based federal regulation of employment practices. "

The target article is precisely what you would expect from validity generalization 'believers', a plethora of over-confident assertion, claims of 'settled science', and a Nelson-like view of the actual research evidence. However, one peer commentary stands head and shoulders above the rest; that of Brink, K.E., & Crenshaw, J.L. (2011) The affronting of the Uniform Guidelines: From propaganda to discourse. *Industrial and Organizational Psychology*, 4, 4, 547-553.

Some excellent argument and references are provided in this article, most especially to:
Biddle, D.A. (2008) Are the Uniform Guidelines outdated? Federal guidelines, professional standards, and validity generalization (VG). *The Industrial-Organizational Psychologist*, 45, 4, 17-23 and
Biddle, D.A. (2011) Should employers rely on local validation studies or validity generalization (VG) to support the use of employment tests in Title VII situations? *Public Personnel Management*, 39, 4, 307-326.

As Biddle (2008), p. 20 states ...
"Even if a test used by an employer "shows up valid" for 100 other positions/ employers, the challenged employer still has the burden for showing the test is job related for their position in question and consistent with business necessity in their context."

The caselaw and logic in (Biddle 2010) is equally impressive, and stands in stark contrast to the broad, exaggerated claims made by McDaniel et al and those who assume Validity Generalization can replace local validity en masse.

Bottom Line:

Validity generalization can come at a price; unnecessary exposure to legal challenge. For many large organizations, there really is no 'safe' alternative to the routine monitoring of their selection practices, in order that evidence-bases are developed as a matter of course for internal evaluation as well as for potential defence against a grievance claim by disaffected employees or unions.

The 3rd line of Defence

Assessments were made of candidates, but were not used in the decision process.

The validity of this defence depends entirely upon whether the assessment results were seen by any member of the selection panel. Why? Because:

1. *if* no member of the assessment panel saw any scores, *or*
2. had any knowledge of how candidates performed or even behaved during the assessments, *and*
3. the assessments were never discussed with any candidate as part of the selection process,

then clearly the assessment scores could not have affected the selection outcomes.

However, if 1, 2, or 3 above is not the case, then it will be difficult to explain how the knowledge was excluded. Just saying "we ignored the scores" is open to adversarial challenge. In reality, the HR selection panel may really have faithfully declined to use the scores for some reason *(that reason is itself important to the court, as that judgement not to use the scores may open up other lines of questioning).* But, the issue is whether seeing the scores for an individual, even if to be discounted, would still have affected the subjective decision-process employed by members of the selection panel. For example, if seeing a very low test score for Conscientiousness for an individual *(which would be indicative of a sloppy and poor attitude to work)*, is it plausible that a decision-maker could simply discount this knowledge entirely from an employment decision?

I think the only effective response here has to be computational.

That is, a defence against a challenge of 'unconscious or conscious bias' revolves around examining the patterns of test scores among candidates, relative to who was finally selected and rejected. What needs to be shown is that selection was effectively random with respect to the patterns of test scores among candidates.

This is not an easy task since the test scores would not be distributed randomly among candidates, but would likely follow a heavily skewed beta-distribution pattern, where few scores are low on any attribute. A re-sampling/ bootstrap approach might be envisaged here where it can be shown that the outcome of the selection process could have occurred quite frequently without any knowledge of the test scores.

Two Intriguing Issues

1 From page 3 ... in response to the 1st line of Defence initial ploy ...

1 Psychometric test scores are like any other kind of information about a candidate. They describe features and attributes of a candidate in ways not described by their job performance, references, supervisor ratings and discussions, and interview-related information. All of this information is carefully discussed and taken into consideration by the selection panel. Sometimes, test scores might well be overridden by cogent arguments from the panel.

The key argument here is that the test scores are like any other kind of information used by a decision-maker. All such information contains error to some extent, and it is the reasoned judgment of the decision-makers which is applied to the mix of information in order to arrive a final decision. The problem with this line of argument is that it only succeeds in moving the evidence-requirement from the test scores to the behaviour of the decision-makers themselves.

And that raises an intriguing question ...
... if test scores are treated as no different from any other information about a candidate, such as biodata, interview data, reference data, supervisor rating data, then why should their validity be given any more scrutiny than to the validity associated with those other information sources?

Put another way, if test or assessment-centre scores are just components of a complex information-rich subjective decision process, in which little or no formal attention is paid to the validity of the other information sources, why do we pay so much attention to the validity of these particular scores?

2 Validity coefficients apply to situations where the test score (or some function constructed from other contributing predictors) is used as the sole predictor of an outcome. When HR or a selection panel claim they incorporate test scores into a mix of information which forms the basis for making final decisions, the validity evidence for a test score no longer holds, because the indicative magnitude of a score can now be overridden by other information the selection panel feels is more relevant to a particular candidate decision.

That subjectivity may enhance the validity of a test, or reduce it.

Given test scores are far from perfect predictors of outcomes, such 'subjective but experienced' augmentation may indeed prove to be more effective than if just selecting candidates on the basis of test scores alone. But it is a matter for empirical evaluation, not speculation.

However, the point remains, a test score validity coefficient assumes that the test score alone is used as the selection device. If you do not use test scores this way, then the validity coefficients you perhaps invoke as 'evidence-bases' may no longer be relevant to your particular selection application.

In Conclusion

The aim of this whitepaper has been to examine a possible employee grievance scenario, and how an HR executive with assistance from the test publisher/consultancy might typically approach a defence of the selection practices and assessments used.

These defences are taken in turn, and shown to potentially fail, or at least become 'questionable' under subject-matter-expert adversarial challenge.

This is not an exhaustive document as any grievance likely contains specific features which will need to be addressed with a customised approach.

However, the brief elaborations and evidence bases put forward here do show that whether or not they were aware of this state of affairs, legal exposure to many kinds of selection practices (not just psychometric tests) is a reality for HR.

The simple solution preferred by US courts, is for a company to validate its own selection procedures in-house. For commercial, ideological, or perhaps statistical reasons, many test publishers and I/O psychology consultants propose a 'validity generalization' defence, which renders local validity obsolete.

However, there are sufficiently powerful counter examples of validity not generalizing, and good reasons explaining why this can happen, to suggest that relying upon validity gathered in other organizations to justify the utility of a particular assessment is not the optimal strategy for a company wishing to avoid legal exposure.

Acquiring local validity and ongoing program evaluation is in fact a reasonably straightforward organizational systems-procedure which should be undertaken for any routine intervention imposed upon a workforce (leadership development, team-building, coaching, selection processes, promotion strategies etc). This is after all a 'critical feature' activity for a truly strategic HR.

References

Barrett, P.T. & Rolland, J.P. (2009). *The meta-analytic correlation between the Big Five personality constructs of Emotional Stability and Conscientiousness: Something is not quite right in the woodshed* [Whitepaper]. Retrieved from http://www.pbarrett.net/stratpapers/metacorr.pdf?cls=file

Biddle, D.A., & Nooren, P.M. (2006). Validity generalization vs Title VII: Can employers successfully defend tests without conducting local validation studies?. *Labor Law Journal*, 57, 4, 216-237.

Biddle, D.A. (2008). Are the Uniform Guidelines outdated? Federal guidelines, professional standards, and validity generalization (VG). *The Industrial-Organizational Psychologist*, 45, 4, 17-23.

Biddle, D.A. (2011). Should employers rely on local validation studies or validity generalization (VG) to support the use of employment tests in Title VII situations? *Public Personnel Management*, 39, 4, 307-326.

Faust, D. (2011). *Coping with psychiatric and psychological testimony 6th Edition*. Oxford, UK. Oxford University Press.

Johnson, J.W., Steel, P., Scherbaum, C.A., Hoffman, C.C., Jeanneret, P.R., & Foster, J. (2010). Validation is like motor oil: Synthetic is better. *Industrial and Organizational Psychology*, 3, 3, 305-328.

McDaniel, M.A., Kepes, S., & Banks, G.C. (2011). The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, 4, 4, 494-514.

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

Schmidt, F.L., & Oh, I. (2010). Can synthetic validity methods achieve discriminant validity? *Industrial and Organizational Psychology*, 3, 3, 344-350.

Steel, P., Johnson, J.W., Jeanneret, P.R., Scherbaum, C.A., Hoffman, C.C., Foster, J. (2010). At sea with synthetic validity. *Industrial and Organizational Psychology*, 3, 3, 371-383.

Webster, C.D., Müller-Isberner, R., & Fransson, G. (2002). Violence risk assessment: using structured clinical guides professionally. *International Journal of Forensic Mental Health*, 1, 2, 185-193.

Webster, C.D., Douglas, K.S., Eaves, D., & Hart, S.D. (1997). *HCR-20: Assessing Risk for Violence (Version 2)*. Simon Fraser University: Vancouver, Canada.

Ziskin, J. (1981). *Coping with psychiatric and psychological testimony 3rd Edition*. Venice: California. Law and Psychology Press.

Ziskin, J. (1995). *Coping with psychiatric and psychological testimony 5th Edition*. Venice: California. Law and Psychology Press.