A Rejoinder to David Bartram, October 2008


Dr Almuth McDowall, University of Surrey

a.mcdowall@surrey.ac.uk

Dr Rainer Kurz, Saville Consulting

rainer.kurz@savilleconsulting.com

Rab McIver

Rab.maciver@savilleconsulting.com

Professor Peter Saville

Peter.saville@savilleconsulting.com

Abstract

This brief article is a direct and invited rejoinder to the comments by David Bartram on our original paper on the use of 360 degree feedback in coaching. Bartram raises important issues which are likely to be of concern and interest to coaches and other users of psychometric tools, even if they do not relate directly to the primary focus of our original paper. This rejoinder reiterates our original approach which was based on a practical workshop on coaching which included a case study reinforcing the benefits of good structural alignment and validity between psychometric trait measures and criterion measures of work effectiveness. This we consider a logical and practical extension of the criterion-centric perspective to measurement, which is a measurement perspective we share with Bartram.

Next, our rejoinder addresses how the Great Eight model mapping was used in the analysis as a unit weighted aggregate to predict overall effectiveness at work rather than predict behavioural competencies. Overall effectiveness (and ability) is central to Saville Consulting Wave® Performance 360 criterion measure, and we put this in the context of a hypothetical gap analysis to make the link back to the application in coaching.
High level models which lead to aggregation of both predictors and criterion variables are increasingly commonplace in the literature and include Bartram (2005). We further this work by the aggregation of overall (global) work effectiveness measures. Rather than differentially weighting these predictors we used a priori unit weighting of the Great Eight predictors from Wave® Professional Styles and OPQ32i. We were primarily concerned with testing the hypothesis that the results were non-zero in predicting overall effectiveness in the original article.

We confirm that Wave® Professional Styles trait measure of Competency Potential clearly improved on OPQ32i's prediction of overall effectiveness (p<.05, two tailed, N=169) in this study.

Finally, we emphasise the advantages of single co-validation studies in the comparison of the validity of different models and psychometric tools based on fully pre-hypothesised equations to aggregate predictor scales. This approach, paired with other methodologies, we argue will lead to the scientific advancement of the field.

Keywords: performance, potential, matched model, criterion domain, trait

**Introduction**

This is a short response to Bartram’s 2008 reply to our original paper on the Effective Integration of 360 Degree Feedback in the Coaching Process. We value the points raised, and are glad to address these point by point.

1) Bartram questioned ‘what role publication of the report of this study played in a paper that was ostensibly about integration of 360 feedback into the coaching process’?

2) Why were the Great Eight computed from OPQ32i and not used in the analyses?

3) Why was a single aggregate predictor computed and then compared with three ostensibly different single item criterion measures?

4) Why did the authors not test the statistical significance of the differences between the correlations found?

1) Bartram found it difficult to see ‘what role publication of the report of this study played in a paper that was ostensibly about integration of 360 feedback into the coaching process’.

The original paper resulted from an invitation by the journal editor to write up two popular skill based sessions at the 2007 Special Group in Coaching Conference at City University. These practical conference contributions featured a case study where behaviour, ability and global (BAG) aspects of work effectiveness on the Saville Consulting Wave® Performance 360 results were compared with Job Profiler data expressed in exactly the same competency language, and with Wave® Styles and Swift aptitude test results. Critically, the interpretation of the case study, which was made available to all participants, centred on the unique Saville Consulting Wave® ‘Matched Model’ that aligns role
requirements, 360 performance measures and predictions from personality or styles measures (i.e. Wave Professional and Focus Styles). This enables various forms of gap analysis:

1. How does performance compare to role requirements?
2. How does potential compare to role requirements?
3. How does potential compare to measures of performance?

Executive coaches address exactly these issues with a multitude of clients in their daily work. They often rely on an intuitive and experience based grasp of the challenges and opportunities facing an individual. Competency based assessments of role requirements and performance can add structure and rigour to the process, while psychometric trait measures help to ‘explain’ poor or outstanding performance, and enable users to gauge how easy or difficult it will be to address a performance issue. The value of these assessments is to a high degree contingent on the tools having been developed in an integrated framework based on a validation-centric approach which maximises validity between psychometric trait measures and performance based (e.g. 360 feedback) or criterion assessments focusing on overall performance as well as behavioural and cognitive components of performance.

Through 360 degree feedback coaches can access the coachee perspective and gain input and buy in from the manager. Self, managers’ and other stakeholders’ ratings of role requirements provide further scope for discussion of current performance, help in the setting of future objectives and expectations and coaching interventions that aim to deliver performance improvement in key areas. The alignment and validity of the predictor (e.g. Wave® Styles) and criterion measures (e.g. Performance 360) is a fundamental concern if both are to be used in conjunction as part of the coaching process as we recommend, and an integral part of our original article. The workshop slides and case study materials are available free-of-charge from the authors. The purpose of our article was to share these thoughts more widely, explaining the Saville Consulting Wave® notion of structurally aligned predictors and criteria and the empirical rationale therefore, before returning to our core purpose which was to outline how 360 degree feedback profiles can be used in coaching. We will reply briefly below to the psychometric issues raised, having invited Professor Peter Saville and Rab McIver to our rejoinder due to their authorship of the OPQ and Wave® model respectively and their continuous involvement with this research.

2) Why were the Great Eight computed from OPQ32i not used in the analysis?

Great Eight measures were computed from OPQ32i as an intermediate step towards an OPQ32 total prediction score.

We referred to this OPQ32 composite score as ‘Corporate Leadership score’ (note that the actual OPQ32 Corporate Leadership report was not referenced and its equations are unpublished to our knowledge) but in the light of the points raised by Bartram acknowledge that it would be better to refer to it as ‘OPQ32 Great Eight Total’ in future. In this study on N = 169 the OPQ32i Great Eight Total correlated .49 with the Wave® Professional Styles Competency Potential Great Eight Total
(significance p<.01 two tailed). This indicates that the overall measures of these two questionnaires are tapping into related if not identical overall broad constructs.

Saville Consulting Wave® Performance 360 was developed from the outset to measure behaviour, ability and global (or overall) work effectiveness. This approach therefore expands from data-driven behaviour-only competency models such as the Great Eight into new domains to create a hierarchical model of performance which encompasses both ability and critically overall effectiveness at work (147 components are available at the most detailed level). In other words, the measurement and diagnoses of both ‘micro-level’ effectiveness in terms of very specific behaviours but also global assessment of performance is crucial to the overall model.

Thus, it was important to show that each of these three different types of criteria (Behaviour, Ability and Global) aligned to measures in the predictor domain as we regard trait (predictors) and performance (criteria) and the integration of these two forms of data as potent information for leveraging behavioural change as part of a wider coaching process.

We illustrate this here with a brief a gap analysis. According to the model, we expect an individual who scores high on the styles or self report personality variable which underpins a competency on Wave® Focus to gain high scores on criterion scores of Performance if the model is well aligned and valid. So if predictor competency potential scores on ‘Creating Innovation’ in the trait measure are high, indicating that the individual feels both able and motivated to innovate, we would also expect that they score highly on criterion measures, here 360 degree feedback scores, that relate to this aspect of performance (if the predictors and criterion are well aligned and valid). In reality however, coachees may exhibit gaps between their styles competency potential prediction (from Wave® Styles) and the criterion evaluation given by raters (from Performance 360). They may for example have an overall low score in the criterion rating of Creating Innovation in contrast to a high self report trait score. The analysis of this gap provides coach and coachee with valuable information and a structure for discussion. One possible hypothesis that the coach may test is that there are situational or personal factors which are hindering the effective deployment of this competency, so that the underpinning competency potential is not being realised. These factors may or may not be under the individual’s control. For instance, it is possible the respective raters do not see the individual as effective in Creating Innovation because the culture at work provides little opportunity to put forward ideas (situational factor). Equally, it is possible that the individual may not be explaining, promoting or delivering their ideas (personal - which may become apparent from their scores on other competencies). Finally, it is possible that there is a difference between how effective ‘innovation’ is understood and therefore rated by different parties (where different people perceive different behaviours as being hallmarks of ‘innovation’). Probing these gaps and the potential reasons for them provides an insightful mechanism to initiate and sustain behavioural change.

That behavioural competencies link to personality and styles measures is not controversial and has been extensively reported in other work (Robertson & Kinder, 1993; Bartram, 2005) and for Performance 360 Behavioural component validation itself has already been outlined in MacIver (2006) as referred to and briefly summarised in our paper. As a result the individual eight variables from of three different predictor scores were not shown against the four different criteria in the interest of space and relevance.
As readers may not be familiar with the ability and global domains covered in Performance 360 our intention in the article was to briefly investigate how these aligned to established predictor instruments and models such as Wave Styles, the Great Eight and OPQ from Bartram 2005.

3) Why was a single aggregate predictor computed and then compared with three ostensibly different single item criterion measures

There is now a trend towards aggregating personality measures to higher order scores (e.g. Digman, 1997), as there is evidence that personality can be understood in more succinct terms than the ‘Big Five’ (Barrick & Mount, 1991) or other more finely grained models. More recently, Musek (2007) has gone further and provided a rationale for a Big One. Such higher level aggregates are important in understanding the performance domain, and understanding the validity that is actually available to the user from reports and profiles (see MacIver, et al., 2008), rather than the validity which theoretically can be extracted from a tool with complex multivariate statistics. How such aggregates are formed is critical to the performance or criterion centric measurement to make available the maximum validity to the test user.

How can we measure performance in highly complex roles? This is the question that sits at the heart of the criterion problem. The key challenge for the profession here is to move from ill-defined single item ‘Overall Performance’ measures to a more discerning understanding of this construct. Nyfield et al. (1995) developed for the IVS (International Validation Study) Proficiency and Promotability scales with six items each but did not publish the content of the items. Bartram (2008) reports a reliability of .70 (presumably Internal Consistency) for that Proficiency scale but provides no indication of the item content either. We separated the Proficiency construct into Expertise and Outcome related themes to create our new constructs. The former theme deals with pre-requisites for success at the work place while the latter deals with work results. We took great care to define each item with reference to multiple levels in the organisation to incorporate contextual aspects of performance. We developed the Promotability scale into a Demonstrating Potential item that captures extra role behaviours on the job that are indicative of the individual’s potential to quickly grow into more demanding roles.

Thus, building on Nyfield we developed three items which cover overall performance concisely through a short scale that is widely applicable. In our study the Internal Consistency scale reliability was .71 (N=169).

Bartram (2005) produces aggregated predictor and criterion measures for each study to support the aggregated operational validity of OPQ with aggregated predictors that he quotes in his reply. So aggregation per se is clearly not the fundamental issue. Is weighting the competencies and predictors better? Adapting and Coping and Supporting and Co-operating Great Eight measures have frequently been negatively weighted in the studies of Bartram (2005). These two competency predictors are underpinned by Emotional Stability (.86) and Agreeableness (.90) factors of the Big Five (Bartram, 2005). The worldwide literature however generally highlights these as positive predictors of overall performance not negative (Tett, Jackson & Rothstein, 1991; Salgado, 1997 & 1998; Hurtz & Donovan, 2000). Negative weighting of competencies in predicting overall performance also seems to run counter to the definition of Robertson, Callinan & Bartram (2002): "We define competencies as sets of behaviours that are instrumental in the delivery of desired
results or outcomes. Bartram’s argument appears to be that we can aggregate, but need to weight the predictors. However, the creation of weighted scores which predict ‘disagreeable neurotics’ as superior performers at work is in direct contradiction to the extant literature. From inspection of our own data base and OPQ normative and ipsative intercorrelation matrices we advance the hypothesis that these negative statistical weights are caused by the ipsative only nature of OPQ used in the vast majority of predictor and criterion measures used by Bartram (2005).

Our approach, which is stable, practical and usable, was to use conservative unit weight equations which we specified prior to analysis rather than relying on a data-driven approach alone. Our approach avoids capitalising on sample specific chance effects or statistically mining the data for the optimal solutions and is in line with the literature on the direction of different trait measures in predicting performance.

4) Why did the authors not test the significance of the differences between the correlations found?

The design of our study set out to test whether one simple equation worked across all criteria. We showed conventional statistical significance tests to establish where the results were non–zero. Tests of statistical significance are only meaningful if the study design carries an appropriate level of statistical power. At the outset of the study modest validity gains were expected that would have required sample sizes well in excess of 169 to have the appropriate level of statistical power hence no statistical test of significance was applied. Comparison of validity of the different effectiveness measures was not part of the study design as it did not relate to the hypotheses being tested.

We included results to show that the Saville Consulting Wave® Performance 360 overall effectiveness measures could be measured using a criterion centric (the Great Eight mapping aggregate Total) approach from two questionnaires (Wave® and OPQ) driven by a criterion centric model. Here, one questionnaire performed better than the other in measuring overall (global) effectiveness at work. The differences between OPQ and Wave® Styles Competency potential on Overall Effectiveness are statistically significant (two-tailed p<.05). This indicates that creating Competency Potential Wave Great Eight predictor scales demonstrated better criterion related validity when unit weighted in assessing overall effectiveness at work, as defined here, than the published OPQ equations (Bartram 2005) when unit weighted. As illustrated by the example given above, this alignment between the predictor and criterion measures allows coach and coachee a forum to discuss potential differences between individual traits and performance through gap analysis.

We therefore feel that the statement that “Wave Professional Styles…. clearly improved on the prediction offered by the well-established OPQ32i tool” is justifiable.

Benefits of Co-validation Driven by Pre-Hypothesised Equations
Bartram points to the limited reliability of single item criterion measures. The logical impact of this reliability is likely to drive down the observed validity of both measures. Bartram does not comment on why the validity coefficients we report are statically significant and have meaningful effect sizes.

Bartram puts forward that a single study on different instruments has shortcomings. All study designs have shortcomings, meta-analyses, for example, are likely to suffer from the file drawer effects to different degrees which it is difficult if not impossible to quantify or control for. We are of course aware that one piece of evidence is indeed just one piece that pertains to a larger jigsaw and validation at its best should be an ongoing process of hypothesis testing (Landy, 1986). We uphold that there are also clear advantages to co-validation using one sample, where a whole host of extraneous variables which can affect the validity are kept constant allowing for a better comparison between the validities of instruments.

A key advantage of our approach was that we did not use multiple regression, or canonical analysis with ipsative data. We investigated straightforward hypotheses that we could set out in advance across different personality instruments to predict overall effectiveness at work.

This direct comparison of instruments and models is much needed in the academic literature, as uncorrected validities allow for a finely grained comparison which is informative to researchers and practitioners alike. A paper outlining the full methodology will be made available in 2009.

We argue that our approach is indeed informed by and builds on Bartram’s seminal work as much as on other research in the field. There are many similarities between Bartram’s and our drive to improve our knowledge of the criterion space and understanding how best to predict it. We see benefits of both substantial co-validation studies and meta-analyses of these studies in the future.

**Conclusion**

In summary, we welcome this opportunity to address the points raised by Bartram. We uphold that valid 360 instruments are a robust tool for use in coaching particularly where the exact assessment and subsequent development and improvement of work performance are the fundamental issues to be addressed. Our original skills sessions and subsequent paper offer a valid model for understanding workplace performance through the matched model approach. Our paper was balanced in that we addressed a) the rationale for using 360 degree feedback in a coaching context, b) the necessity to understand how performance can be understood and measured c) offered co-validation evidence which provides support for the Wave® model and d) offered the discussion of an actual profile.

Bartram contended in his reply that we had failed to make clear the core purpose of our article and failed to support our claims for validation with appropriate psychometric evidence. We have argued above that the direct comparison of validity coefficients in uncorrected single studies provides a valuable method for gathering support for a priori specified models which should form the basis of subsequent meta-analytic techniques in the future.
As such, we are cognisant that there is a need for both approaches in order to generate more evidence for and further our understanding of the best criterion model of workplace performance, and its respective predictors. We sincerely hope that in due course more researchers will follow our path and publish their own co-validation studies.

Acknowledgement

We would like to thank an anonymous reviewer for constructive comments on the draft submission.

References


