Running head: Two-stage alignment

Let’s Focus On Two-stage Alignment, Not Just On Overall Performance

Dave Bartram,
SHL Group Ltd, UK

Peter Warr,
Institute of Work Psychology, University of Sheffield, UK

Anna Brown
SHL Group Ltd, UK

Commentary article on Johnson et al, “Validation is Like Motor Oil: Synthetic is Better”

Industrial and Organizational Psychology: Perspectives on Science and Practice.

Johnson et al. (2010) note the need for a vector defining the relationships between job components and overall job performance (OJP) in the job requirements matrix approach to synthetic validation. This need is also implicit in the job components validity approach which they discuss. In our view relationships between job components and overall criteria like OJP are central to synthetic validation, as the effectiveness of this approach depends on the generalizability of relations between job component-criterion relationships across jobs, organizations and (of increasing importance) countries.

Within that perspective, we believe that both approaches to synthetic validation described in the paper suffer from an over-reliance on OJP as the ultimate criterion. This has consequences which pose problems for any validation approach, whether synthetic or otherwise. While both the job components and job requirements approaches acknowledge the role of job elements, those components are seen as intervening variables, rather like a means to an end and not of intrinsic importance themselves.

Our preference is to shift the focus of interest from ill-defined global constructs like overall job performance (OJP) to specific elements of a job. It is important to focus on components of workplace behavior, since relationships between predictors and components are expected to be generalizable across situations; in contrast, aggregates and other composite criterion measures are considerably more variable in their nature and correlates.

We propose a two-stage model in which we differentiate two classes of criteria: measures of performance on specific job components on the one hand and more general measures, like OJP, on the other. We can closely align ‘predictors’ (such as personality scales, ability measures and other assessments used in selection) with job components. However, it is not possible to align these same predictors with less well defined measures like OJP. OJP can be partly 'explained' by a weighted combination of job component measures, together with other situationally dependent factors, however the weights given to job components will tend to vary from situation to situation. The relationships between predictors and criterion components on the other hand are more generalizable.

Problems with OJP and Other Variably-defined Overall Criteria

Overall performance (OJP), promotability and similar gross constructs are insensitive indicators and should not be used as the basis for judging the validity of instruments or the generalizability of validity. As multi-component constructs they are amalgams of more specific behaviors, sometimes of very different kinds. An individual may perform well in one respect (e.g., dependable work) but be less effective in other component activities (say, innovative behavior), and in some cases component behaviors can have negative associations with each other (e.g. aspects of supporting and cooperating competencies on the one hand and task-motivated contributions on the other). An overall index, seeking to represent diverse subordinate themes through a single value, is necessarily imprecise unless between-component homogeneity is great.

A second problem with overall job performance (OJP) derives from the fact that the behaviors which are particularly valued (and thus make up indices of overall performance) differ from organization to organization and from job to job. For instance, attention to detail and a critical approach are central to overall performance in quality control work, whereas in customer support jobs a polite and friendly manner is more essential. Bartram (2005) examined the associations in different studies between each of eight behavioral competencies and ratings of overall job performance. The behaviors contributed to OJP in different ways between the studies; for any one of them, associations with OJP ranged in different studies from around zero to strongly positive (see Table 16 in that paper).

Predictions of overall job performance (supposedly a single construct) can thus have a criterion variable which differs between organizations and between jobs. Each study's overall performance is determined by its own assortment of constituents and their relative importance in that particular job setting. As a result, different studies may examine criterion constructs that are dissimilar despite giving them all the same label as overall performance. Furthermore, the meta-analytic accumulation of OJP findings from different studies frequently places together non-equivalent overall criterion variables which have different constituent behaviors and which thus represent different constructs.

A third complication is that the several behaviors that make up overall performance are themselves predictable from a particular trait in different ways. Individual associations between a trait and key behaviors can substantially diverge from each other, perhaps being both positive and negative for different elements of overall performance. However, those associations (usually unmeasured) become diffused within the single correlation between that trait and overall job performance. For example, a trait like 'outspoken' may be positively correlated with a competency like 'Initiating action' and negatively with 'Team cohesion'. The two competencies in turn may have positive correlations with OJP. In such a case, 'outspoken' would not appear to be a predictor of OJP.

In examining the criterion-related validity of a personality measure it is therefore inappropriate to ask without qualification whether it significantly predicts overall job performance. A strong association between a trait and overall job performance is expected only in particular circumstances: when that trait is substantially correlated with behaviors which are themselves major constituents of the overall indicator. The validation issue for a personality inventory thus becomes: are the traits being measured likely to predict behaviors which are identified as central to overall performance in a particular setting? A significant trait-OJP association is not expected in the absence of trait-behavior-overall alignment. It follows that overall performance without two-stage matching is inappropriate as a criterion for examining a measure's validity.

For these several reasons, research literature indicating that personality scales and factors are typically unrelated to overall performance (e.g., Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt. 2007a, b) is misleading. Published validation studies in respect of overall criteria fail to examine patterns in terms of constituent behaviors and trait-related hypotheses about those behaviors in particular settings.

The Need for Conceptual Matching Between Predictors and Criteria

While there is general acceptance of the view that validation research should be based on conceptual matching between predictors and criteria (e.g., Campbell, 1990; Hogan & Roberts, 1996; Robertson & Kinder, 1993; Schneider, Hough, & Dunnette, 1996; Tett, Jackson, & Rothstein, 1991), only a few publications have compared associations when a predictor and criterion are or are not thematically matched (Bartram, 2005; Hogan & Holland, 2003). Warr (1999) analyzed 480 correlations between 30 personality traits and 16 rated job behaviors as a function of the conceptual concordance between each trait and each behavior. As expected, the average trait-behavior correlation in the absence of conceptual overlap was about zero (-0.02), and it was progressively larger as conceptual concordance became greater; for high trait-behavior concordance the average correlation was 0.25.

We are currently preparing a report on five studies that make it clear that personality can be a valid predictor of an overall performance indicator when that overall indicator is substantially composed of behaviors that are themselves predicted by the traits examined. The findings indicate that overall indicators are inappropriate as validation criteria unless the predictors studied are likely to be linked to behaviors of that kind. While this may seem self-evident, studies continue to be published reporting low 'validities' for predictors where inappropriate or poorly measured criteria are used. We therefore argue that a two-stage focus is essential, going beyond the conventional over-simple examination of trait-OJP correlations without regard for key intervening behaviors in a particular setting. A two-stage approach has long been established as good practice in personnel selection, but its more careful application in validation research is still required.

Bartram (2005) used the term 'criterion-centric' to argue for a shift in focus away from predictors and towards what we need to predict. Indeed, the whole notion of predictor instruments 'having' criterion-related validity is misleading. Because of inherent variability in the form and nature of measures like OJP, the apparent validity of a predictor can change from low to high from study to study due solely to variation in the properties of the criterion and its unmeasured constituent behaviors. It is far more appropriate to attribute validity coefficients to the relationships between specific predictor-criterion pairs (where the criteria are at the job component level of measurement). In those cases, estimates are likely to more stable and conceptually justifiable.

Even in this case we know that the validity of a predictor-criterion relationship, using the same instrument as a predictor and the same people as criterion raters, can vary substantially depending on the method of measurement used for the criteria. Bartram (2007) compared the validities obtained using either forced-choice or Likert scales as the basis for line manager ratings of competencies, with the line managers using both methods to rate the same set of job incumbents. For the same predictor instrument (OPQ), operational validities were 0.38 for the forced-choice criterion measures and 0.25 for the Likert ratings. Thus, simply by changing the format of the criterion measurement instrument one can produce a 50% increase in the apparent validity of a predictor instrument. This entails that any attempt to generate synthetic validity estimates for novel situations will be affected on the ways in which criterion job components were measured in those studies that feed into the component validity estimates.

Discriminant Validity

This observation relates to the concern raised by Johnson et al. (2010) regarding the poor discriminant validity associated with predictor-criterion relationships. Improving discriminant validity is important for a synthetic validity approach, as otherwise any predictor might be expected to predict any criterion. We believe an essential prior step is the development of an appropriate well-articulated model for the job behavior domain. As put forward in Bartram (2005), the key to this lies in identifying relatively uncorrelated areas of behavior (labeled as the “Great Eight”) and then specifying separate predictor scales in terms of which of those areas they most relate to. Using this approach, Bartram reported very good discriminant validity: the average correlation between unmatched predictor-criterion pairs was -0.02 while that for matched pairs was 0.16 (uncorrected for artifacts). Add to that reliable procedures for designing criterion measures that force raters to discriminate (Bartram, 2007), and we may well be able to raise substantially the levels of discriminant validity associated with component level relationships and thus demonstrate real utility for the synthetic validity approach.

Can we agree on 'universals'?

We strongly support the principle of synthetic validity and agree with much of the focal paper, seeing that form of validity as the only general approach for the future. However, we have to be very clear about differences in the natures of the constructs we are using. At least three logically distinct types of measure must be considered: measures of potential (lead measures), measures of behavior (current measures) and measures of achievement or performance (lag measures). The first two of these can, at least in principle, be well-specified and be generalizable across situations. The third by its nature cannot. The third category of measure (achievement or performance) is necessarily to some degree situationally, culturally or organizationally specific. It also has the greatest measurement problems associated with it, as the constructs people are asked to assess are generally ill-defined, complex, have a large time-span and are multidimensional. Linked to that, we have argued for the investigation of two-stage alignment rather than the traditional emphasis on overall performance indicators alone.

If we are to pay more attention to the components intervening between predictors and overall performance measures, it is important to reach agreement on the level of aggregation or disaggregation in the criterion space. We have advocated the "Great Eight" competencies as providing a good level of aggregation for optimizing discriminant validity. In addition, the SHL

Universal Competency Framework has more specific levels, including 20 competency dimensions and 112 competency components (Bartram, 2005). In some cases, it is appropriate to work at that more specific level to obtain better discrimination between jobs and roles. However, this level of specificity does raise challenges for obtaining criterion measures that are themselves adequately differentiated.

We suspect that agreement on SHL's model or any other as *the* universal framework would be difficult to achieve. As a consequence any system for aggregating data from diverse contributors into a single database would need to provide the possibility of representing that data in terms of a number of different construct systems.

Is a Shared Approach to Synthetic Validity Possible?

Johnson et al (2010) argue for the development of a shared database to underpin the development of synthetic validity. The practical constraints on implementing a shared database lie in large part in the commercial interests to be dealt with. As a commercial company, SHL has a large database of detailed job analyses linked through its Universal Competency Framework to the requirements of particular jobs from which we can generate synthetic predictions of scales' validity in respect of new jobs. Naturally this is built around our own models and predictors, and the framework represents a significant financial and intellectual investment. Others have similar investments to consider. However, before reaching the stage of resolving the practical and commercial issues relating to how such data sources might be merged to the benefit of all, we need to develop consensus on the constructs that should be measured and quality standards for the measures obtained – especially with regards to the conceptualization and measurement of criteria at several levels of specificity.

Commercial arrangements for pooling data could be worked out on the basis of data trading business models, whereby companies contributing content to the system received tradable 'credits' that could then be used to pay for usage of the system. By placing a charge on use of the system by non contributors, any surplus of credits could be turned into cash by net contributing companies. However, such an arrangement would only make sense if it was developed on a sufficient scale to make it commercially viable and if it embodied sufficient flexibility to accommodate different job component construct models.

Conclusions

We have suggested that the primary current limitation on progress is coming from the criterion measurement space and not from the predictor space. Well-designed, reliable instruments are available that can measure specific aspects of people's potential with high levels of construct validity. However, at present we do not have comparably-high standards of measurement in the criterion space. We argue that a model of two-stage alignment is key to resolving this, as OJP-type criteria are intrinsically ill-defined. However, this in turn entails the need to agree on the constructs such an intervening component model should embody or agreeing on a range of models that can be mapped onto each other.

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185-1203.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*, 263-272.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology*, vol. 1, pp.687-732. Palo Alto, CA: Consulting Psychologists Press.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology*, *88*, 100-112.

Hogan, J., & Roberts, B.W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior*, *17*, 627-637.

Johnson, J. W., Steel, P., Scherbaum, C. A., Hoffman, C. H., Jeanneret, P.R., & Foster, J. (2010). Validation is like motor oil: Synthetic is better. *Industrial and Organizational Psychology: Perspectives on Science and Practice*

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683-729.

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology*, *60*, 1029-1049.

Robertson, I. T., Baron, H., Gibbons, P., Maciver, R., & Nyfield, G. (2000). Conscientiousness and managerial performance. *Journal of Occupational and Organizational Psychology*, *73*, 171-180.

Robertson, I.T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology*, *66*, 225-244.

Schneider, R.J., Hough, L.M., & Dunnette, M.D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior*, *17*, 639-655.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Warr, P.B. (1999). Logical and judgmental moderators of the criterion-related validity of personality scales. *Journal of Occupational and Organizational Psychology*, *72*, 187-204.

Warr, P.B., Brown, A., & Bartram, D. (in preparation). Overall performance, promotability and the validation of personality measures.