*Industrial and Organizational Psychology*, **3** (2010), 344–350.
Copyright © 2010 Society for Industrial and Organizational Psychology. 1754-9426/10

# Can Synthetic Validity Methods Achieve Discriminant Validity?

FRANK L. SCHMIDT
*University of Iowa*

IN-SUE OH
*Virginia Commonwealth University*

Our focus is on the difficulties that synthetic validity encounters in attempting to achieve discriminant validity and the implications of these difficulties. Johnson et al. (2010) acknowledge the potential problems involved in attaining discriminant validity in synthetic validity. For example, they report that Peterson et al. (2001), Johnson (2007), and other synthetic validity studies report failure to achieve discriminant validity. What this failure means is that a synthetic validity equation developed to predict validity for Job A does as well in predicting validity for Jobs B, C, D, and so forth as it does for Job A. Johnson et al. then go on to propose that this problem might be overcome by careful attention to both the criterion and predictor sides of synthetic validity. We question whether their proposals can be made to work.

Correspondence concerning this article should be addressed to Frank L. Schmidt.
E-mail: frank-schmidt@uiowa.edu

Address: Department of Management and Organizations, Tippie College of Business, University of Iowa, W236 John Pappajohn Business Building, University of Iowa, IA 52242-1994

Frank L. Schmidt, Department of Management and Organizations, Tippie College of Business, University of Iowa; In-Sue Oh, Department of Management, School of Business, Virginia Commonwealth University.

## Criterion Considerations in Discriminant Validity

On the criterion side, the major obstacle to attainment of discriminant validity is halo error in ratings of job performance. In the vast majority of studies, performance is measured by ratings, usually supervisory ratings. If raters are not capable of discriminating between different dimensions of job performance, then it will not be possible for synthetic validity methods to differentially predict performance on these separate dimensions, thus ruling out the possibility of discriminant validity. In their discussion of halo error, Johnson et al. discuss only the Viswesvaran, Schmidt, and Ones (2005) study, which found high levels of halo error in both supervisory and peer ratings of job performance dimensions. They state that this study may have ''blurred'' dimensions of job performance when rating dimensions from primary studies in the meta-analysis were assigned to the job performance dimension categories used in their meta-analysis, thereby producing an inflated estimate of halo error in ratings. Johnson et al. do not mention any of the many other studies of halo error. Thirty years ago, King, Hunter, and Schmidt (1980) reanalyzed ratings data from 11 such primary studies, published mostly in the 1970s. They found that on average across these primary studies, 31% of the variance



https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press

in observed ratings was due to halo error (the percentage for the true scores on the ratings was even higher: 50%). This finding means that on average the correlation between observed ratings from any given rater and halo error is .56 (and .71 for the true scores). As they point out, this is a large amount of halo error. It is comparable to the halo finding in Viswesvaran et al. for observed ratings, and in the studies reviewed by King et al. there was no assignment to job performance dimension categories that could have caused a ''blurring effect.'' In contrast with halo error, job performance dimensions accounted for an average of only 8% of the variance. Mount, Judge, Scullen, Sytsma, and Hezlett (1998), in a large sample study, also found that rater effects (halo error) were much larger than the effects of job performance dimensions. Again, there was no possibility of a ''blurring effect.'' In short, there is a large and long-standing research literature calibrating halo error in ratings, and there is agreement across these studies that halo error is quite large. In rating a given ratee, raters appear to discriminate among performance dimensions to only a very slight degree. Their ratings on all performance dimensions are dominated by their global impression of the ratee's overall performance (Viswesvaran et al.).

Johnson et al. state, ''Rater training and rating format can reduce halo error.'' But as King et al. (1980) note, past research shows that changing the rating methods, or instructions, or specificity of the performance dimensions has not been successful in reducing halo. In fact, the King et al. study tested the hypothesis that forced choice rating scales would reduce halo error; they found little support for the hypothesis. Failed attempts to reduce halo error go back many years. In the late 1960s, when the first author was a PhD student at Purdue, Hubert Brogden stated in his classes that the Army Research Institute in Washington (which he headed) had tried many different procedures to reduce halo in ratings, including rater training, but that none of them worked. He reported that even the psychologists at the Institute who studied halo had a lot of halo error in their own ratings of the performance of others! The implication of all this is that the discriminant validity in job performance ratings that is required by synthetic validity methods is highly unlikely to be realized in practice.

Johnson et al. state that ratings of different dimensions of performance can in fact be differentially predicted, and they cite Dudley, Orvis, Lebiecki, and Cortina (2006) as demonstrating that this can be done. The relevant results purportedly showing this are presented in Table 4 of Dudley et al. We examined the (small) variation in mean validity estimates across the four dimensions of job performance (task performance, job dedication, interpersonal facilitation, and counterproductive work behaviors [reflected]). We found that all this variation is accounted for by second-order sampling error, using the methods presented in Hunter and Schmidt (2004). There are no nonartifactual differences in validity across the performance dimensions and therefore no true differential prediction of these performance dimensions—as would be expected from what research has shown about halo error.

## Predictor Considerations in Discriminant Validity

Based on a very large military data set, Schmidt, Hunter, and Pearlman (1981) reported that the validity of any given cognitive aptitude test was essentially identical across jobs that varied widely in their task makeup. They concluded that these findings indicated that any given aptitude test (e.g., verbal aptitude) or general mental ability (GMA) test will have about the same validity for all jobs, so long as these jobs are at the same level of complexity. Johnson et al. responded to this finding by challenging the broader finding in the research literature that job complexity moderates the validity of cognitive tests. They state that this belief is mostly based on a technical report by Hunter (1983) in which he reordered the

https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press

complexity classifications to fit the pattern of validities. In fact, Hunter never reordered the three basic data categories he used. He reordered only a small job category (i.e., one with few incumbents in the U.S. labor force) that was defined by very high standing on the things dimension of the Data, People, and Things (DOT) system. In any event, the finding that job complexity moderates validity has been repeatedly independently replicated. Salgado, Anderson, Moscoso, Bertua, and de Fruyt (2003) replicated this finding based on European data, showing that it generalizes across countries and cultures. Gutenberg, Arvey, Osburn, and Jeanneret (1983) also found that the mental complexity dimension of jobs moderated test validity. The Bertua, Anderson, and Salgado (2005) meta-analysis is another study supporting this conclusion. The moderating effect of job complexity is well supported in the literature and is certainly not dependent on a single study or data set.

Johnson et al. cite Verive and McDaniel (1996) as a study that failed to find that job complexity moderates the validity of GMA. However, Verive and McDaniel did not examine any measure of GMA. They examined only short-term memory, which is not a cognitive aptitude construct, much less a measure of GMA. McDaniel (Michael McDaniel, personal communication, January 25, 2010) stated that this study cannot be interpreted as a test of whether complexity moderates the validity of GMA or any cognitive test. Johnson et al. also cite Hulsheger, Maier, and Stumpp (2007) as evidence against the moderator role of job complexity. The authors of this study state that the lower mean validities they found for high-complexity jobs are due to the peculiarities in the German educational system, which causes applicant pools for high-complexity jobs to be much more homogenous on GMA than in the case of the United States and other European countries. They describe the effect of this educational system as ''indirect range restriction'' on applicant pools and state (p. 12), ''As indirect range restriction has a repressing effect on operational validities, correcting for indirect range restriction would certainly raise operational validities to the level found in U.S. or European meta-analyses.''

Finally, Johnson et al. cite Steel and Kammeyer-Mueller (2009) as contradicting the complexity moderator finding. But the conclusion in this study is based on an unusual analysis. Typically one would classify jobs by complexity level and then look to see whether mean validity for a given test increased as complexity level increased; this is the approach used in past studies. However, Steel and Kammeyer-Mueller used a regression equation to predict *individual observed validity coefficients* from the complexity levels of the jobs. When they found that this multiple correlation fell just short of statistical significance ($p = .11$), they concluded that job complexity did not moderate validity. There are two things wrong with this procedure: (a) It relies on a statistical significance test and assumes that failure to attain statistical significance means there is no relationship when in fact it may be due to low power (Schmidt & Hunter, 1996); and (b) it employs an extremely unreliable criterion: individual validity coefficients. We know from validity generalization studies that most of the variance in individual observed validity coefficients is due to sampling error and other statistical and measurement artifacts. The most accurate estimates are that such artifacts account for between 82% and 87% of the variance in observed validities (Schmidt et al., 1993; Table 3). This means that reliability of a set of validity coefficients, taken as a criterion measure, is somewhere between .18 and .13. (Note that sampling error and other artifactual variance functions here as measurement error.) A criterion with such low reliability is very difficult to predict because all correlations with it are greatly attenuated by its unreliability. Only small correlations with it are possible. Statistical power to detect such small correlations is low. And any statistically significant relation found may be due to chance or capitalization on chance.

https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press

## Prediction of Validities by PAQ Dimensions

Johnson et al. acknowledge that in past synthetic validity studies, attempts to predict test validities from the PAQ information and similar job analysis measures have not been very successful. Multiple *R*s have been low. First, they argue that the reason for this was that past studies have used ordinary least squares (OLS) multiple regression. They propose that use of weighted least squares (WLS) multiple regression (in which validities are weighted by the inverse of their sampling error or sample size) results in substantial ability to predict validities from PAQ dimensions. However, this weighting system has already been built into the percentage variance accounted for figures cited above (82–87%); hence, the reliabilities are .13 to .18. This means that use of WLS regression (though better than OLS regression) does not avoid the problem of low criterion reliability and validity; the reliability index sets the upper limit of validity.

Second, Johnson et al. state that use of WLS regression with PAQ dimensions (13 or 32 dimensions) as predictors allows substantially higher prediction of validities than that with DOT dimensions (only three dimensions), based on Steel and Kammeyer-Mueller (2009). A more likely explanation for this may be capitalization on chance (Cattin, 1980) that inflates multiple *R*s more as the number of predictors increases. Failure to correct for this inflation may account for their finding that the multiple *R* was smallest when the 3 DOT dimensions were used, larger when the 13 PAQ dimensions were used, and largest of all when all 32 PAQ dimensions were used. This latter multiple *R* was .63 ($R^2 = .40$; p. 543). Psychometrically, it is not possible that a criterion with a reliability in the range of .13 to .18 could be predicted with an unbiased multiple *R* of .63. (Steel and Kammeyer-Mueller used validities corrected for range restriction and criterion unreliability; such corrections do not increase the reliability and significance level of the validities used as a criterion because the corrections also increase the standard error; Hunter & Schmidt, 2004, Ch. 3.) It may well be that these three multiple *R*s would be shown to have very similar or identical (small) values when the proper shrinkage formula (Cattin, 1980) is applied.

## Further Considerations in Discriminant Validity

Even if job complexity did not moderate validity of cognitive tests, that fact would not bolster the synthetic validity argument by Johnson et al. That is, even if job complexity were not a moderator, it would still be the case, as shown by Schmidt et al. (1981), that jobs that are very different in task makeup show virtually identical levels of validity for any given cognitive test. This finding strongly suggests that a task analysis of jobs cannot be used to predict the validity of cognitive tests because validity of cognitive tests is not related to the task makeup of jobs. A fundamental assumption of synthetic validity methods is that analysis of the task content of jobs can be used to predict the validity of different tests for those jobs. For cognitive tests, this assumption is contraindicated by research findings.

Johnson et al. cite Peterson et al. (2001) and Johnson (2007) as finding that synthetic validity equations do not have discriminant validity; that is, an equation developed for Job A will predict validity just as well for Job B (or any other job) as for Job A. This finding confirms the conclusion that task analysis of jobs cannot differentially predict validities. That is, the careful task analysis and other time consuming and expensive steps required in the synthetic validity process have essentially no effect on the final validity estimates and therefore have no effect on the final synthetic validity battery chosen to predict job performance. This casts suspicion on the notion that synthetic validity methods are solidly based and scientific. Critics will argue that because the complex synthetic validity process has little or no effect on the final validity estimates or on the final tests chosen for use, the

https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press

process is more akin to Voodoo than to science.

These considerations mean that for any given cognitive test synthetic validity equations must predict about the same level of validity for most or all jobs. This explains the finding described in Johnson et al. that synthetic validity estimates for cognitive abilities are fairly accurate when compared against validity generalization estimates or estimates from large sample individual studies. This level of accuracy is explained by the fact that validity generalization and large sample single-study validity estimates for cognitive abilities vary little across jobs, and this is true for synthetic validity estimates too. Hence, the synthetic validity estimates appear to be accurate. (A reviewer stated that this means that validity generalization, like synthetic validity, lacks discriminant validity. But validity generalization does not claim or require discriminant validity across jobs. Its main finding is generalization across jobs not discrimination between jobs; unlike synthetic validity, it is not based on assumptions of discriminant validity between jobs.)

Most of the studies cited by Johnson et al. in support of the accuracy of synthetic validity estimates compare these estimates to observed validity coefficients and conclude that synthetic validity estimates are accurate if they fall within the 95% confidence interval around the observed validity coefficient (Hoffman, Holden, & Gale, 2000; Morris, Hoffman, & Shultz, 2003; Peterson et al., 2001). These confidence intervals are often quite wide. By this test, synthetic validities are accurate. But the discriminant validity question is this: Can the synthetic validity equations *differentially* predict validity across jobs? In the three studies cited here, the correlation between the synthetic predicted validities and the empirical observed validities were low, indicating a lack of discriminant validity. This is the explanation for the finding, discussed above, that synthetic validity equations developed for one job work just as well for other jobs.

Johnson et al. cite one study (Hoffman & McPhail, 1998) that they say found a high correlation between synthetic validity estimates and independent empirical estimates of the same validities. However, an examination of this study reveals that its results cannot be generalized to individual jobs because the synthetic validity estimates used were averages across 51 different jobs. As the authors explain, ''We chose to calculate mean predicted JCV (Job Component Validities) in this study as a way to control for errors in PAQ and JCV analyses at the level of the individual job'' (p. 995). They also state, ''This averaging provided a more stable estimate of validity for the JCV procedure than would be possible in studies which examine a single job'' (p. 999). The authors state that this averaging across 51 different jobs is one of two important reasons their study ''provided substantially greater convergence between JCV estimates and empirical estimates'' than previous studies (p. 999). The problem here is that synthetic validity is supposed to be able to be applied to individual jobs. Thus, the results of this study are not really relevant to synthetic validity as it is intended to be used in real applications to particular jobs.

The empirical validity estimates used in the Hoffman and McPhail (1998) study were the mean observed validity generalization estimates from Pearlman, Schmidt, and Hunter (1980). Each mean was based on many individual studies (up to 882 studies). This validity coefficient is much more reliable than the individual observed validities used in the other synthetic validity studies, and this is the second reason for the larger relationship between the mean JCV values and the validity criterion. However, as we have seen, Johnson et al.'s claim is that synthetic validity can predict the validities obtained from individual empirical validity studies not that it can predict average validities across many studies.

Johnson et al. acknowledge the research finding that specific aptitudes (such as verbal, quantitative, and spatial) contribute little to incremental validity over and above a reliable measure of GMA. They state that the abilities domain is an ''interesting

https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press

exception'' to the general rule that the best prediction is obtained when narrow criteria are matched with narrow predictor measures, and they cite research evidence showing that narrower measures of personality can sometimes predict narrow criteria better than broader personality measures. They hypothesize that ''expanding the predictor space to include a wider variety of predictors than has typically been studied in synthetic validation research (e.g., personality, biodata, situational judgment) . . . should improve our chances'' of attaining discriminant validity on the predictor side.

How likely is it that this hope can be realized? It is well known that the best predictor of job performance (and training performance) is GMA (or combinations of specific aptitudes that are essentially identical to GMA measures; cf. Hunter, Schmidt, & Le, 2006). Therefore, synthetic validity equations, if they are accurate, must predict a high level of validity for GMA for all jobs. Because the goal in synthetic validity is not just to produce a selection system with *some* validity but rather to provide maximal validity (and therefore maximal utility or practical value), all the resulting synthetic validity predictor batteries must include a GMA measure or its equivalent. Omission of GMA measures would mean suboptimal levels of validity. GMA measures have much higher validity than personality, biodata, or situational judgment tests. For example, based on a large amount of cumulative data, Schmidt, Shaffer, and Oh (2008) showed that the mean validity of GMA tests is nearly three times (2.91 times) as large as that of Conscientiousness measures and 5.52 times as large as that of Emotional Stability measures. Because of this, GMA will virtually always dominate the predictor composite produced by synthetic validity equations or batteries for predicting job performance (i.e., GMA will have the largest weight in the equation or battery). This dominance of synthetic validity battery composites by GMA or equivalent measures will make it nearly impossible for these equations or batteries to have discriminant validity, even when they contain noncognitive predictors. A synthetic validity equation or battery developed to predict performance on any given job will predict performance about equally well for many other jobs, jobs very different in task make up and in other ways.

## Conclusion

Although we are favorable to the idea of synthetic validity as a general concept, we believe that the research evidence indicates that the synthetic validity methods described in Johnson et al. cannot achieve meaningful levels of discriminant validity. Johnson et al. agree that discriminant validity is a critical requirement if synthetic validity is to have a sound basis and be credible. In light of the research findings summarized in our comment, we do not believe that this requirement can be met. The prospects for discriminant validity are bleak both on the criterion side and on the predictor side.

## References

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A U.K. meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387–409.

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65*, 401–414.

Dudley, N., Orvis, K., Lebiecki, J., & Cortina, J. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57.

Gutenberg, R. L., Arvey, R. D., Osburn, H. G., & Jeanneret. P. R. (1983). The moderating effects of decision-making/information processing job dimensions on test validities. *Journal of Applied Psychology, 68*, 602–608.

Hoffman, C. C., Holden, L. M., & Gale, K. (2000). So many jobs, so little ''*N*'': Applying expanded validation models to support generalization of cognitive test validity. *Personnel Psychology, 53*, 955–991.

Hoffman, C. C., & McPhail, S. M. (1998). Exploring options for supporting test use in situations precluding local validation. *Personnel Psychology, 51*, 987–1003.

Hulsheger, U. R., Maier, G. W., & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment, 15*, 3–18.

https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press

Hunter, J. E. (1983). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Department of Labor, Employment Service.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.) Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594–612.

Johnson, J. W. (2007). Synthetic validity: A technique of use (finally). In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 122–158). San Francisco: Jossey-Bass.

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

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced choice performance evaluation scale. *Journal of Applied Psychology, 65*, 507–516.

Morris, D. C., Hoffman, C. C., & Shultz, K. S. (2003, April). *A comparison of job components validity estimates to meta-analytic validity estimates*. Poster presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology, 51*, 557–576.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373–406.

Peterson, N. G., Wise, L. L., Arabian, J., & Hoffman, R. G. (2001). Synthetic validation and validity generalization: When empirical validation is not possible. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits of personnel selection and classification* (pp. 411–451). Mahwah, NJ: Erlbaum.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*, 573–605.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task difference and validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66*, 166–185.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78*, 3–13.

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy of range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61*, 827–868.

Steel, P., & Kammeyer-Mueller, J. (2009). Using a meta-analytic perspective to enhance job component validation. *Personnel Psychology, 62*, 533–552.

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23*, 15–32.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131.

https://doi.org/10.1111/j.1754-9434.2010.01249.x Published online by Cambridge University Press