Industrial and Organizational Psychology, **3** (2010), 371–383.
Copyright © 2010 Society for Industrial and Organizational Psychology. 1754-9426/10

RESPONSE

At Sea With Synthetic Validity

PIERS STEEL
University of Calgary

JEFF W. JOHNSON
Personnel Decisions Research Institutes

P. RICHARD JEANNERET
Valtera Corporation

CHARLES A. SCHERBAUM
Baruch College

CALVIN C. HOFFMAN
Los Angeles County Sheriff's Department and Alliant University

JEFF FOSTER
Hogan Assessment Systems

Abstract

We expected that the commentary process would provide valuable feedback to improve our ideas and identify potential obstacles, and we were not disappointed. The commentaries were generally in agreement that synthetic validity is a good idea, although we also received a fair amount of suggestions for improvements, conditional or tempered praise, and explicitly critical comments. We address the concerns that were raised and conclude that we should move forward with developing a large-scale synthetic validity database, incorporating the suggestions of some of the commentators.

In 1957, researchers representing industry, government, and academia came together in Chicago for a symposium on synthetic validity. The contents of this symposium were published almost intact 2 years later in *Personnel Psychology*. In 2009, we came together for a similar symposium in New Orleans. Taking inspiration from what happened 52 years ago, we collaborated to put forth this focal article on synthetic validity. Our goals were to summarize how far we as a field have come, to address challenges, and to inspire us to go the remaining distance. How did we do? Let's look to the commentaries.

On the positive side, there seems to be general agreement that synthetic validity is a good idea. Bartram, Warr, and Brown (2010) write, ''We strongly support the principle of synthetic validity and agree with much of the focal article, seeing that form of validity as the only general approach for the future.'' Hollweg (2010) thinks ''a database that promotes this goal would be

Correspondence concerning this article should be addressed to Piers Steel.
E-mail: piers.steel@haskayne.ucalgary.ca
Address: Piers Steel, Haskayne School of Business, University of Calgary, SH444 - 2500 University Drive, N.W. Calgary, Alberta, Canada T2N 1N4; Jeff W. Johnson, Personnel Decisions Research Institutes; Charles A. Scherbaum, Department of Psychology, Baruch College; Calvin C. Hoffman, Los Angeles County Sheriff's Department and Alliant University; P. Richard Jeanneret, Valtera Corporation; Jeff Foster, Hogan Assessment Systems.

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

a worthy achievement and add significantly to the scientific body of the profession.'' McCloy, Putka, and Gibby (2010) argue that substantial funds for realizing it will materialize ''given the potential value to the field.'' Murphy (2010) admits ''there is no doubt that synthetic validity is a great idea.'' Oswald and Hough (2010) are ''convinced that the payoff [for building a synthetic validity database] would directly benefit the welfare of organizations (in real dollars) as well as employees and the science of work behavior.'' Russell (2010) strongly agrees that ''synthetic validity is 'practically useful' and . . . that it also holds value in developing theory.'' Schmidt and Oh (2010), despite several pointed criticisms, ultimately indicate that they are ''favorable to the idea of synthetic validity.'' Finally, Vancouver (2010) ''advocates for the development of a synthetic validity database along the lines described by Johnson et al. (2010), not only because of its practical utility but also because of its potential contribution to the science of I-O psychology generally.''

Almost unanimously, synthetic validity is highly praised, with the majority of commentators believing that this is the future for our field. Not to be misleading, we also received a fair amount of suggestions for improvements, conditional or tempered praise, and explicitly critical comments. We sorted these comments into six categories. The first two categories address discriminant validity concerns, specifically (a) ''are you sure you can discriminate validity between jobs?'' and (b) ''are you sure you can discriminate between performance dimensions?'' The third category is ''Sure it's better, but is it better enough?'' This is mostly a reply to Murphy (2010), who questions whether synthetic validity is significantly better than validity generalization, engaging us to make sure that the added effort needed to create synthetic validity is worth the result. The fourth category, ''the personal touch,'' addresses how expert opinion fits into the synthetic validity process. Fifth, ''what else can it do'' considers other applications and research questions that synthetic validity can address. Finally, we end with ''been there, done that.'' In this section, we recognize that there are many near-synthetic validity projects and smaller synthetic validity projects currently underway. Their number underscores how feasible a large-scale synthetic validity project can be once we have developed the proper tools.

Following these six sections, we outline next steps. Borrowing Murphy's (2010) metaphor that synthetic validity is a ship that has already sailed, we argue that synthetic validity has been successfully sailing the selection seas for decades, stopping into port periodically for retrofitting. Synthetic validity is ready for another upgrade, so let's lay out the gangplank. As the consensus of opinion revealed here, we have plenty of people who want to come onboard and bring synthetic validity to its next destination.

Are You Sure You Can Discriminate Validity Between Jobs?

The most ardent criticisms of synthetic validity come from Schmidt and Oh's (2010) commentary. Frank Schmidt along with Jack Hunter must be given recognition for pioneering some of the basic work that gave rise to the modern synthetic validity framework. Notably, their work disputing the then widespread notion that validities are situationally specific made modern synthetic validity possible. Our field is deeply indebted to their efforts, as ''this was among the most significant changes in organizational research over the past century because it allowed knowledge to accumulate much more rapidly across studies'' (Steel & Kammeyer-Mueller, 2008, p. 55). However, their viewpoint reflects an opposite extreme that there is no substantive situational specificity for validity coefficients and what little remains can be accounted for by general level of complexity.

For job component validity (JCV) to work, there must be predictable variance in validity coefficients across different jobs. Consequently, we agree with Schmidt and Oh's (2010) basic idea that job complexity

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

does moderate validity coefficients. We also agree with Schmidt and Oh that validity coefficients are not easy to predict, especially because of measurement error, and that Hunter's (1983) DOT complexity scale will reach significance if reexamined with far more cases. However, this statement can be made of almost any predictor (e.g., Meehl's crud factor), and it is not the focus of the issue. We want to know the *best* way to operationalize job complexity so that it predicts those hard-to-predict validity coefficients, not just *if* job complexity can predict. This requires us to determine if O*NET or PAQ dimensions are superior to those derived from the DOT. How can we determine which is better?

Schmidt and Oh (2010) are correct that ''typically one would classify jobs by complexity level and then look to see whether mean validity for a given test increased as complexity increased.'' However, the typical procedure is not necessarily correct. Comparing means without consideration of within-group variance is a fundamentally flawed procedure, with Harvey (2010) reminding us about the problems of predicting central tendency while ignoring variability. Means alone do not tell the percentage of variance accounted for, and this is what we are interested in (Lubinski & Humphreys, 1996). Weighted least-squares regression provides the requisite statistic, a recommended technique for meta-analytic data if accompanied with outlier analysis (Hunter & Schmidt, 2004; Steel & Kammeyer-Mueller, 2002). To this end, Steel and Kammeyer-Mueller (2009) showed that the PAQ predicts approximately 30% of the variance of general mental ability (GMA) validity coefficients, with the DOT data dimension predicting about 1%. Schmidt and Oh hypothesize that the PAQ does better because of overfitting because it uses more predictors. There is no need for conjecture on this point. Using Alf and Graf's (2002) maximum likelihood correction, the R^2 shrinks from .30 to .25. Steel and Kammeyer-Mueller also did an exploratory estimate based on 32 subdimensions, acknowledging that a larger database is needed to give a more accurate estimate. Corrected, this R^2 shrinks from .40 to .28. Accounting for about 25 times more variance, the PAQ is a better way to operationalize job complexity than the DOT data dimension. More importantly, job complexity does effectively predict validity coefficients.

Finally, Schmidt and Oh (2010) note that Steel and Kammeyer-Mueller's (2009) results are questionable because Schmidt and colleagues have estimated that between 82% and 87% of observed variability among validity coefficients is attributable to measurement error. In conjunction with what was predicted by the PAQ, this means that between 107% and 115% of variance is accounted for, which is impossible. However, it is also impossible to give a generic percentage of variance that is because of measurement error. Exactly how much measurement error exists in a meta-analytic database is primarily determined by sample size (Murphy, 2000), which will vary from case to case but can be readily computed. In Steel and Kammeyer-Mueller's analysis, sampling error comprises 55% of the variance for GMA validity coefficients, whereas in the article relied on by Schmidt and Oh for their estimates, the comparable estimate is 81% (Schmidt et al., 1993). Also, Schmidt et al. used a more homogenous sample, sorting by job families and then averaging results, which will further reduce variance predictable by differences between jobs. In other words, they removed the exact variance we are trying to predict. As Murphy (2000) concludes, ''a frequent criticism of validity generalization [meta-analysis] procedures is that they are biased in favor of the hypothesis that validity is consistent'' (p. 199). Perhaps this viewpoint has some merit. There is plenty of variability left for JCV to predict validity coefficients after taking into account the measurement error.

Given that there is relevant variance to account for and that we have already successfully done so, the subsequent question is how to proceed. Should we use the O*NET or the PAQ? Vancouver (2010) makes the sensible conclusion of perhaps

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

neither as ''additional constructs should be added.'' So far, our endeavors to predict validity coefficients have primarily focused on those involving GMA, and it is not clear whether similar successes would occur with validity coefficients based on other predictors. Consequently, we should consider what we have achieved here as a good start but not the finished product. One avenue to explore is offered by McCloy et al. (2010) as well as Oswald and Hough (2010), who suggest examining several O*NET factors, such as work context scales, as moderators. Also, Vancouver suggested we consider time on the job or experience as a moderator as well, although he was concerned that it would make the project too onerous. We do not believe this to be a problem. As Steel, Huffcutt, and Kammeyer-Mueller (2006) conclude, we can add time-varying covariates later as the database expands to capture this variance. Although modeling time on the job would improve synthetic validity, it does not have to be immediately pursued because it isn't a necessity. For synthetic validity to be viable, it doesn't have to account for all the variance among validity coefficients, just a useful amount of variance and this is a threshold we have already passed.

Are You Sure You Can Discriminate Validity Between Performance Dimensions?

Although it is not necessary for job component validation, an implicit assumption of the job requirements matrix approach to synthetic validation is that we can differentiate among performance dimensions. We do not see discriminant validity among performance dimensions as a ''critical requirement if synthetic validity is to have a sound basis and be credible'' (Schmidt & Oh, 2010), but it is certainly a desirable quality. The job requirements matrix approach is a construct-oriented approach to personnel selection, using job analysis information to link relevant predictors to specific job components. This not only allows for the demonstration of validity evidence in jobs that are too small for a traditional criterion-related validation study but also ensures that the selection procedure is job related, which should enhance legal defensibility and user acceptance. Although Schmidt and Oh warn us that ''critics'' will argue that this process is ''more akin to Voodoo than to science,'' we believe this is a highly scientific approach to personnel selection (cf. Schneider, Hough, & Dunnette, 1996) regardless of whether an alternative weighting of the predictors provides a nearly equal validity coefficient.

At this time, there is too little research to draw definitive conclusions about the discriminant validity of synthetic validity equations. The degree to which a validity coefficient for a job-specific equation differs from the validity coefficient generated by another job's equation depends on the type of predictors available and the similarity of the jobs. If jobs are similar enough that the same set of predictors can be used to predict performance for all of them, we should not be surprised that the weighting scheme used for one job does not provide very different results from the weighting scheme used for another job. It has been consistently demonstrated that different weights applied to the same predictors often produce similar composite scores and thus similar validities (Koopman, 1988; Ree, Carretta, & Earles, 1998). Research on the discriminant validity of synthetic validity equations that we cited in our focal article has been limited to applying different weights to the same predictors. On the other hand, a synthetic validation strategy allows for the selection of *different* predictors for *different* jobs, so a true test of discriminant validity would be comparing equations that include different predictors. For example, if one equation includes a mechanical comprehension test and another replaces it with a language usage test, we are more likely to see discriminant validity when applying the equation for one job to the other job.

A more fundamental concern is that Schmidt and Oh (2010) contend that we cannot differentiate between job components in the first place. For example, King,

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

Hunter, and Schmidt (1980) found that job performance dimensions accounted for an average of 8% of the variance in observed ratings. This is the traditional viewpoint on the topic, but it is not the final word as a new understanding of this issue is emerging. Previous studies, including many of those cited by Schmidt and Oh, made the assumption that ''if it ain't trait it must be method'' (Lance, Baranik, Lau, & Scharlau, 2009, p. 339). This assumption is not necessarily true. Raters may differ in their performance ratings, but all may be still rating performance, just based on different standards of performance or based on different experiences with the ratee (e.g., a person acts differently with colleagues than around the boss). More recent research has demonstrated empirically that it is more appropriate to interpret rater and source effects as sources of performance true-score-related variance rather than error (Lance, Dawson, Birkelbach, & Hoffman, 2010). Using a more sophisticated approach to model such true-score idiosyncratic rater effects, Hoffman, Lance, Bynum, and Gentry (2010) found results ''consistent with three primary models of performance ratings, *each specifying the impact of empirically distinguishable dimensional performance on ratings*'' [italics added] (p. 144). Again, as Bartram et al. (2010) point out, even more differentiation should occur with motivated raters who are appropriately trained and who are using a well-designed performance appraisal measure.

Having established that it is indeed possible to differentiate between performance dimensions, the next move is to determine what dimensions should be used in a large-scale synthetic validity database. To this end, Bartram et al. (2010) closely consider the benefits of assessing specific performance dimensions instead of just overall performance. We wholeheartedly agree with their assessment, which provides additional reasons for using specific performance dimensions beyond the fact that Job Requirements Matrix (JRM) methodology needs them by its very nature. Furthermore, we find their two-stage alignment solution very compatible with what we proposed (i.e., most performance ratings would be collected at a specific level and then classified into broader categories). Still, the basic problem remains: What will these dimensions be? We are not sure. In truth, our scientific field is still determining what aspects of performance we are capable of distinguishing and under what conditions. As Bartram et al. conclude, ''the primary limitation on progress is coming from the criterion space and not from the predictor space.'' Addressing this issue, we suggested starting out with broad performance categories and then going to more specific categories as more data are accumulated and letting the data determine when sufficient specificity has been reached. Bartram et al. also usefully point out that there will be no single answer to this issue as several different taxonomies will likely be effective, including SHL's Great Eight and O*NET's Broad Nine. Ideally, a finished synthetic validity system will allow users to choose among several configurations of performance. Developing an appropriately flexible performance taxonomy is our first order of business.

Sure It's Better, But Is It Better Enough?

Murphy (2010) provides one of the harder issues to address. It is not a question of can we build a synthetic validity system, which he believes we can, but whether we should. Although synthetic validation provides significantly better results than most local validation studies (in terms of lower standard errors around the validity coefficient, smaller sample sizes necessary within jobs, and greater flexibility for estimating validity in jobs not included in the original study), how does it compare to validity generalization? Validity generalization and synthetic validity are like brothers, related techniques that are based on similar assumptions (Jeanneret, 1992). Although we would expect some sibling rivalry, it appears we must determine which is to be our favorite son.

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

The case for validity generalization is that it provides a simpler and immediately available way of devising a selection system—simply use something like GMA and Conscientiousness on the basis of their mean validity coefficients across primary studies. This is not a bad idea as both constructs are typically good predictors of performance and they both are universally valid (i.e., credibility intervals do not cross zero) because jobs everywhere benefit from having smart and hardworking employees. For many reasons, however, we do not see validity generalization to be superior to synthetic validation.

First, the fact that a test is universally valid just means its validity coefficient will always be on one side of zero. To properly design and evaluate a selection system, however, we need to know the size of the validity coefficients. It tells us how much to weigh predictors relative to each other, and it is vital for utility analysis, to determine how much we should invest in the entire selection process. We also need precise validity coefficients to calculate the likelihood of a candidate meeting a cutoff, both for minimum and maximal performance. The Productivity Measurement and Enhancement system, for example, builds around the insight that higher levels of performance are not always significantly more valuable, that there are diminishing returns (Pritchard, Paquin, DeCuir, McCormick, & Bly, 2002).

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–.42, with even larger ranges for its facets (Dudley, Orvis, Lebiecki, & Cortina, 2006). Again, this represents plenty of practical variability that should be predictable using synthetic validation.[1]

Furthermore, once you go beyond GMA and Conscientiousness, predictors tend not to generalize in dependable ways. Murphy (2009a) gets around this problem by asserting that this is all we really need because most other predictors are derivable from these two (i.e., positive manifold). Later, Murphy (2009b) admits that this is more true for the cognitive than the personality domain, highlighting that there are still several relevant predictors that cannot be justified through validity generalization alone (Tett & Christiansen, 2007). Still, Murphy is quite right that the remaining predictors almost certainly will not be as important as GMA and Conscientiousness, and that we have encountered a law of diminishing returns. However, we interpret this situation as emphasizing the need for synthetic validity. The selection field is beginning to stagnate as there are no more easy pickings left and it is not cost-effective for any single organization to pursue what remains. We contend that there are dozens of small, incremental steps left to be taken, such as evaluating personality at a facet level (Rothstein & Goffin, 2006; Schneider et al., 1996; Tett & Christiansen, 2007), each accounting for just a few percentage points of incremental predictive variance but collectively making a substantive difference. In most cases, there will not be a large-enough sample size to recognize they exist, but as we mentioned in our focal article, this is exactly the type

1. Mathematically, the techniques used to establish validity generalization systematically underestimate variance, as Steel and Kammeyer-Mueller (2008) review. For example, we should adjust meta-analytic variance estimates by $K/(K-1)$ but almost never do; it corrects for a downward bias caused by using the observed instead of true mean during calculations. For example, a meta-analysis based on 20 studies (K) will underestimate the true amount of variance by a little over 5%. The larger the meta-analyses, the less this will be a concern, but it is fair to point out that these credibility intervals, broad as they are, are still to some degree underestimates.

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

of advantage that can be accrued through technical economies by using synthetic validation. Sticking entirely to validity generalization, we are left with a field that can only appreciate the largest and most obvious of predictors: GMA and Conscientiousness.

Compounding the matter, so far we have only dealt with the credibility intervals for predicting performance, the primary but not the only criterion of synthetic validity. The synthetic validation methodology can be used to match individuals with jobs (person–job fit), occupations or vocations (person–vocational fit), teams or groups (person–group fit), and organizations (person–organization fit; Kristof-Brown, Zimmerman, & Johnson, 2006). To assess all these different types of fit, we are interested not only in performance of individuals but how satisfied they are in the job. Like performance, job satisfaction is largely a function of individual differences, the work context, and their interaction, all of which can be assessed during a synthetic validation project. Simply by knowing details about individuals and their potential work situation, we should be able to predict whether an occupation or an organization is a good match in terms of *both* performance and job satisfaction. Unfortunately, predictors of job satisfaction do not generalize quite as readily as those for performance (Griffeth, Hom, & Gaertner, 2000; Judge, Heller, & Mount, 2002). They largely depend on a fit between *particular* individuals and *particular* environments.

Even if all these previous issues did not occur, validity generalization has one more major limitation, one that Murphy (2009a) himself brings up: It is not acceptable. As he concludes, ''tests that are manifestly relevant to the jobs in question are likely to be more acceptable to organizations and other stakeholders (e.g., unions and political groups) than tests that are not linked in any clear way (other than their empirical validity) to the job'' (p. 460). Because synthetic validity customizes the results to specific situations based upon a job analysis that heavily involves the stakeholders, it can largely avoid this weakness. We specifically discussed this advantage in the focal article when we evaluated synthetic validity's legal standing.

Validity generalization does have a significant advantage over synthetic validation in that validity generalization can be applied to small jobs in the absence of an appropriate synthetic validity database. This is a practical advantage that we hope to remedy through the development of the large-scale synthetic validity database that we have advocated. Conceptually speaking, however, we believe that synthetic validity is superior to validity generalization. It is not just that synthetic validity is the older brother, preceding validity generalization in its development. Rather synthetic validity has all the advantages of validity generalization plus the ability to further predict validity coefficients through effective moderator searches. If we want the science of I-O psychology to move beyond ''general cognitive ability and Conscientiousness predict overall job performance,'' we need to look at more specific relationships.

The Personal Touch

Synthetic validity represents a mechanical, or more aptly, an algorithmic system. We let the numbers do the talking. Although we are happy when such math optimizes our Internet searches, Hollweg (2010) argues that it might work against synthetic validity by causing it to be used as a reductionist robotic formula. He worries that a widely available synthetic validity database could produce an overreliance on a mechanized system that underemphasizes the importance of expert judgment within our field. We share this concern to some extent, but we believe that marketing, education, and possibly limiting accessibility to qualified professionals are tools that can do a great deal to alleviate this concern. The purpose of the synthetic validity system would be to facilitate expert judgment and, if used appropriately, would allow practitioners to explore multiple possibilities when creating a selection system for a client. Certainly, overall validity is not the only criterion

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

to consider, but output from this system could help practitioners weigh other concerns such as costs and accessibility along with more readily accessible and accurate validity estimates.

Hollweg (2010) raised other concerns about how the database could hurt practitioners. Although these concerns are valid, we do not envision a future in which reputable I-O consultants have been replaced by a synthetic validity database. Take the advent of validity generalization as an example of a huge technical advance. We have known for many years now that general cognitive ability is a good predictor of overall job performance for just about any job, but the employment test publishing industry is much stronger now than it was before validity generalization became mainstream. What we are proposing is really just the next logical step, establishing validity generalization for specific predictors and specific criteria. It is important to remember that the public database would not link specific tests to job components but specific predictor constructs. People still need measures of those constructs, and in this age of high-tech cheating, new measures of the same constructs need to be continuously developed. Publishers will continue to try to develop new predictor constructs delivered in new formats in an attempt to do something new and provide value beyond what is available in the database. This may even spur on the next great technological advance. Consultants may wind up marketing against the database by demonstrating through their own research that their instruments are superior to the average level of validity that would be documented for those constructs by the database.

Also advocating a personal touch are McCloy et al. (2010), who suggest an alternative path for developing the synthetic validity system. Instead of empirically deriving correlations, they can be estimated based on groups of subject matter experts (SMEs). McCloy et al. review several steps that would help maximize the accuracy of these estimates, but many of us still have doubts about the underlying mechanism, which is opinion. For example, Harvey (2010) characterized ''subjective 'expert judgment' based methods anchoring the lower end of the scale'' that represents ''a continuum of convincingness.'' Still, McCloy et al.'s technique has given us inspiration about how to improve our own. There is a ''wisdom of crowds,'' and we believe that SME-derived coefficients are likely to reflect reality with some degree of accuracy. We just need to determine to what degree. Consequently, once we have developed a basic but empty synthetic validity matrix, establishing only what the predictors and job components are, we would engage SIOP experts to start filling it out. As we collect empirical data, we can compare these SME-derived coefficients to those observed. If SME-derived coefficients prove to be at least as accurate as what would be derived from a typical local validation study, we could start using the synthetic validity system immediately. Eventually, this preliminary SME-derived matrix will evolve into an empirically derived one as we continue to gather data and swap out coefficients. Meanwhile, we still have a working system.

What Else It Can Do

Other commentaries focused on what else we could do with synthetic validity, suggesting that we were not taking into consideration its full potential. They are right. A lot of features can be incorporated within a synthetic validity platform. In particular, Oswald and Hough (2010) make a series of suggestions. We could explore how to best administer a validation study to maximize correlations. For example, who should make the performance ratings and how? How do differences between job analysis protocols affect results? How can we reduce adverse impact, and what are the tradeoffs with validity? We also are not limited to assessing just individual performance but can consider how the individual would impact team or group performance

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

as well as expand into retention, recruitment, and job satisfaction criteria. Different criteria also open the door to a different set of predictors (Kristof-Brown et al., 2006). As we mentioned and as Harvey (2010) reiterates, synthetic validity is not just for selection. Instead of selecting people for jobs, you can reverse the process and select jobs for people, that is, career guidance and occupational exploration.

Harvey (2010) also makes some other point-on suggestions, although he precedes it with a critique of mean-based JCV. Again, we view mean-based JCV as the ''Plan B'' solution to synthetic validity after early attempts at validity coefficient-based JCV were unsuccessful. As we reviewed in our focal article, recent research has revitalized validity coefficient-based JCV, and we are back to ''Plan A,'' which addresses many of Harvey's concerns. The remaining issues that Harvey brings up are all related to the idea that the linear approach that dominates selection today is far too simple. We agree, especially because these limitations apply most strongly to traditional criterion-related studies. Vancouver (2010) brings up the exact same point that ''we have tended to assume linear relationships between IDs and performance.'' As Vancouver suggests, these interaction effects and nonlinear relationships can be explicitly incorporated during a large synthetic validity project (i.e., they are just additional columns of predictors). Synthetic validity is a way for us to start collecting and codifying these insights as well as immediately distributing the benefits across the selection field once they are confirmed. We are going to be making improvements and fine tuning the process for a long while yet. Without synthetic validity, as we stress, we rely on local validity studies, which typically will not have the power to notice anything but the most pedestrian of findings.

Been There, Done That

Russell (2010) asserts that synthetic validity is much more prevalent than we suggest, as he already possesses or has seen many synthetic validity studies.[2] We are aware that synthetic validity research is out there, but these are generally proprietary technical reports and there has been very little published research on synthetic validity in its nearly 60-year history. Thus, we stand by our contention that synthetic validity is underutilized. In addition, although we have not viewed all the studies to which Russell refers and might readily be proven wrong, what we have seen indicates they are not quite synthetic validity. Rather, they are transportability databases, which are a collection of jobs all with previously developed selection systems. This type of database is used to find a job reasonably similar to a target job and then ''transport'' the validity of the selection system for the similar job to the target job. Transportability, like validity generalization, is related to synthetic validity, as all three are strategies for generalizing validity evidence (Society for Industrial and Organizational Psychology, 2003). Because transportability is not quite synthetic validity, it has its own challenges as Gibson and Caplinger (2007) review. For example, for transportability all the problems associated with small sample sizes which occurred when originally developing any selection system are also transported to each new application. Also, determining if two jobs are adequately or relevantly similar is largely a subjective exercise.

On the other hand, the fact that there are large transportability databases underscores how close we are to building a ''fully loaded'' synthetic validity system, with all the features we described included. These transportability databases can be converted into synthetic validity as soon as the selection systems they are composed of are

2. Of note, we also agree with Russell (2010) that there should be a better analogy than synthetic time, especially because synthetic falsely connotes an ersatz reproduction here. Unfortunately, this is the analogy Lawshe (1952) stuck us with, as per ''The concept is similar to that involved when the time study engineer establishes standard times for new operations'' (p. 32). This is why Lawshe coined the term ''synthetic validity.''

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

developed with the appropriate infrastructure. As we stress, we need to measure performance with a standard taxonomy (i.e., for JRM) and for each job to be assessed with an appropriate job/context/organization analysis measure (i.e., for JCV). This is a good example of our suggestion that ''the creation of synthetic validity can 'piggyback' on what we already profitably do: create selection systems.'' One caveat though: As Russell (2010) indicates, these private databases seem to have focused on a few job families and do not seem generalizable across the entire employment landscape. Still, once properly outfitted, they are a good start and will eventually expand.

The Next Destination

We are very grateful to the commentary authors for showing interest in our focal article and taking the time to share their encouragement, ideas, and concerns. This commentary process has sharpened our own thinking, and we are sure that it has helped clarify issues for the reader. Our primary goal was to publicize synthetic validation techniques as preferred methods for documenting validity evidence for selection procedures, and we feel confident that we have achieved that goal. Another goal was to start the wheels turning on the development of a large-scale, widely available synthetic validity database to advance the science and practice of I-O psychology. We expected that the commentary process would provide valuable feedback to improve our ideas and identify potential obstacles, and we were not disappointed. Although we do not expect smooth sailing ahead, we do not see any icebergs.

We still contend that SIOP is the best institution to initiate the synthetic validity database we have in mind. Similarly, McCloy et al. (2010) believe that ''SIOP would be an ideal host,'' although Hollweg (2010) suggests that we could do better, arguing a hypothetical institution modeled more closely after the Cochrane Collaboration would be preferable. In all, SIOP appears to be the best existing institution for this project. Consequently, we recommend that SIOP initiate a steering committee to bring this concept to fruition. Members of the committee should come from all relevant stakeholders, including representatives from academia, business, and government.

Originally, we did not expect that we could obtain the necessary support from government officials but that has changed. Right now, America is experiencing an almost unprecedented extended period of greatly elevated unemployment (Peck, 2010). The costs of this deep recession are severe and multifaceted, both for the nation and its citizens. Synthetic validity has the potential to help. The job market, like all markets, becomes faster and more efficient if you can improve the speed and quality of relevant information. Adding a synthetic validity component to O*NET seems like a logical step. O*NET already has the taxonomies of predictor constructs, work dimensions, and potential moderators, as well as job analysis data for hundreds of occupations in the U.S. economy. With the data-collection experience gained through the development of O*NET, it may not be as difficult as we imagine to collect predictor and criterion data on selected constructs from incumbents in a wide variety of occupations.

In addition to the economic benefits of improved incremental validity in selection, synthetic validity can help select jobs for people instead of just people for jobs. Used in conjunction with O*NET, which already has a vocational guidance component (Converse, Oswald, Gillespie, Field, & Bizot, 2004), a synthetic validity system that includes the opportunity to self-administer tests measuring various predictor constructs can more accurately identify individual strengths, point people toward occupations in demand, estimate the likelihood that they would be good at them, and help determine whether they would enjoy them. People would also quickly learn what training they need to initiate a career change. Such a sophisticated system that helps people self-select into ideal careers would have a noticeable effect on national GDP, with

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

best estimates in the several hundred *billion* dollar range (Hunter & Schmidt, 1982; Steel & Kammeyer-Mueller, 2006).

Consequently, in return for governmental sponsorship, we provide a very healthy return on investment as we help to maximize national productivity. Pursuing this path allows us to gather the selection data systematically and according to a uniformly high quality. This would be harder to ensure if we took the commercial partnership route, a path that Bartram et al. (2010) pointed out is difficult to realize. Also, government sponsorship would allow a longer and more sustained research effort. Although we can quickly get a working synthetic validity system off the ground, especially if we allow some SME estimation as per McCloy et al. (2010), we would want to continue refining and optimizing this system (Oswald & Hough, 2010). Our ability to improve will continue as we explore, for example, not only the best combination of constructs to predict performance but how each construct should be best operationalized.

Historically, governments have funded exactly this sort of venture to tackle unemployment (Primoff & Fine, 1988). The Social Science Research Council, the National Research Council, and the Occupational Research Program of the U. S. Employment Service all initiated similar efforts. In particular, the U.S. Civil Service Commission, now the Office of Personnel Management, was involved in a massive project involving over 100,000 employees back in 1919. Notably, this is not only a larger number of participants than needed to build synthetic validity today, it was done during a time when it was considerably harder to get them.

The selection industry will also benefit from this initiative in two important ways. First, they need to equip their transportability databases with a few refinements to enable JCV and JRM. Having the appropriate methodology, performance taxonomies, and job analysis for synthetic validity is a necessary outcome of this project. If test developers and selection houses choose to develop their own proprietary synthetic validity systems, using the tools from the database will make it much easier for them. Second, if users want to use this public synthetic validity system for personnel selection, they would need to show that the tests they are using are essentially equivalent to the tests from the synthetic validity system. Such a practice will avoid validity degradation, as would inevitably happen for any single widely used selection battery. Also, such a straightforward validation process opens up what is known as the ''long tail.'' Instead of marketing selection exclusively to large companies, all companies become viable customers. This is especially true for small companies, which have long been expected to benefit from the use of synthetic validity.

We have established we can build synthetic validity, even demonstrating it with working prototypes. Many agree that the field would hugely benefit from a national synthetic validity database. Gary Becker, who won the Nobel Prize for economics, believes, ''in a modern economy, human capital [the work people do] is by far the most important form of capital in creating wealth and growth'' (Wheelan, 2002, p. 106). Human capital and making the most of it is what synthetic validity is all about. This brings up one final question: What are we waiting for?

References

Alf, E. F., Jr., & Graf, R. G. (2002). A new maximum likelihood estimator of the population squared multiple correlation. *Journal of Educational and Behavioral Statistics*, *27*, 223–235.

Bartram, D., Warr, P., & Brown, A. (2010). Let's focus on two-stage alignment not just on overall performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*, 335–339.

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

Converse, P. D., Oswald, F. L., Gillespie, M. A., Field, K. A., & Bizot, E. B. (2004). Matching individuals to occupations using abilities and the O*NET. *Personnel Psychology*, *57*, 451–487.

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.

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

Gibson, W., & Caplinger, J. (2007). Transportation of validation results. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 29–81). San Francisco: John Wiley & Sons.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*, 463–488.

Harvey, R. J. (2010). Motor oil or snake oil: synthetic validity is a tool not a panacea. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 351–355.

Hoffman, B., Lance, C. E., Bynum, D., & Gentry, W. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119–151.

Hollweg, L. (2010). Synthetic oil is better for whom? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 363–365.

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. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette, & E. A. Fleishman (Eds.), *Human performance and productivity: Vol. 1. Human capability assessment* (pp. 223–272). Beverly Hills, CA: Sage.

Hülsheger, 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.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

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

Jeanneret, P. R. (1992). Applications of job component/synthetic validity to construct validity. *Human Performance, 5*, 81–96.

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: Perspectives on Science and Practice, 3*, 305–328.

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*, 530–541. doi: 10.1037//0021-9010.87.3.530.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced choice performance and evaluation scale. *Journal of Applied Psychology, 65*, 502–516.

Koopman, R. (1988). On the sensitivity of a composite to its weights. *Psychometrika, 53*, 547–552.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2006). Perceived applicant fit: Distinguishing between recruiters' perceptions of person-job and person-organization fit. *Personnel Psychology, 53*, 643–671.

Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (2009). If it ain't trait it must be method: (Mis)application of the multitrait-multimethod methodology in organizational research. In C. E. Lance, & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in organizational and social research* (pp. 339–362). New York: Routledge.

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods*. doi: 10.1177/1094428109352528.

Lawshe, C. H. (1952). Employee selection. *Personnel Psychology, 6*, 31–34.

Lubinski, D., & Humphreys, L. (1996). Seeing the forest from the trees: When predicting the behavior or status of groups, correlate means. *Psychology, Public Policy, and Law, 2*, 363–376.

McCloy, R. A., Putka, D. J., & Gibby, R. E. (2010). Developing an online synthetic validation tool. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 366–370.

Murphy, K. R. (2000). Impact of assessments of validity generalization and situational specificity on the science and practice of personnel selection. *International Journal of Selection and Assessment, 8*, 194–206.

Murphy, K. R. (2009a). Content validation is useful for many things, but validity isn't one of them. *Perspectives on Science and Practice, 2*, 453–464.

Murphy, K. R. (2009b). Is content-related validation useful for in validating selection tests? *Perspectives on Science and Practice, 2*, 517–526.

Murphy, K. R. (2010). Synthetic validity: A great idea whose time never came. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 356–359.

Oswald, F. L., & Hough, L. M. (2010). How synthetic validation contributes to personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 329–334.

Peck, D. (2010, March). How a jobless era will transform America. *Atlantic Monthly, 305*, 42–56.

Primoff, E., & Fine, S. (1988). A history of job analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry and government* (pp. 14–29). Toronto: John Wiley & Sons.

Pritchard, R. D., Paquin, A. R., DeCuir, A. D., McCormick, M. J., & Bly, P. R. (2002). Measuring and improving organizational productivity: An overview of ProMes, the Productivity Measurement and Enhancement system. In R. D. Pritchard, H. Holling, F. Lammers, & B. D. Clark (Eds.), *Improving organizational performance with the Productivity Measurement and Enhancement system: An international collaboration* (pp. 3–50). Huntington, NY: Nova Science.

Ree, M., Carretta, T., & Earles, J. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1*, 407–420.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180.

Russell, C. J. (2010). Better at what? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 340–343.

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press

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

Schmidt, F. L., & Oh, I.-S. (2010). Can synthetic validity methods achieve discriminant validity? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*, 344–350.

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.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures*. Bowling Green, OH: SIOP.

Steel, P., & Kammeyer-Mueller, J. (2002). Comparing meta-analytic moderator search techniques under realistic conditions. *Journal of Applied Psychology*, *87*, 96–111.

Steel, P., & Kammeyer-Mueller, J. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods*, *11*, 54–78.

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

Steel, P., Huffcutt, A., & Kammeyer-Muller, J. (2006). From the work one knows the worker: A systematic review of the challenges, solutions, and steps to creating synthetic validity. *International Journal of Selection and Assessment*, *14*, 16–36.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology*, *60*, 967–993.

Vancouver, J. B. (2010). Improving I-O science through synthetic validity. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*, 360–362.

Wheelan, C. (2002). *Naked economics: Undressing the dismal science*. New York. W. W. Norton.

https://doi.org/10.1111/j.1754-9434.2010.01255.x Published online by Cambridge University Press