Industrial and Organizational Psychology, **3** (2010), 340–343.

Better at What?

CRAIG J. RUSSELL
University of Oklahoma

In reading the title of Johnson et al.'s (2010) article on synthetic validity, I immediately asked myself ''Better at what?'' Motor oil serves three functions in internal combustion engines: lubricate, cool, and clean engine parts. There is little doubt that synthetic motor oil does all three of these better than nonsynthetic motor oil, although some might argue that it is not as cost effective. Johnson et al. described two approaches to synthetic validity and then argued why synthetic validity is ''the best approach for many situations.'' I strongly agree with their contention that synthetic validity is ''practically useful'' and with their less directly stated contention that it also holds value in developing theory. Hence, I will limit my comments exactly to how synthetic validity inferences might best contribute to the complimentary goals of advancing theory and practice, although these might cause Johnson et al. to rethink some of their observations.

Proprietary Databases and Synthetic Validity Studies

First, Johnson et al. said ''little coverage in the literature'' resulted in its ''infrequent use.'' Not so, as a large literature of unpublished, proprietary technical reports exists, which relies heavily on synthetic validity inferences (I have seen ~50). I have requested the senior officers responsible for some of these reports to share them with Johnson and one or more of his coauthors (and I have every reason to expect at least some of them to be forthcoming). Centralized databases designed explicitly for transporting criterion validity in personnel selection systems do exist in large consulting firms. These tend to be based on proprietary job analysis processes, although some of these have been published in peer-reviewed journals (Hunt, 1996). With the advent of Internet-based employment testing, these synthetic validity databases rapidly became larger than every meta-analytic total sample I have seen reported.

Most of these predictor–criterion observations were obtained in high volume or high turnover jobs (e.g., call-center positions, retail sales, etc.) with applicants obtained primarily from external labor markets,[1] where consulting firms generate much if not most of their personnel selection system revenue. The job evaluation literature from compensation administration estimates that only about 20% of a typical firm's jobs have external labor markets (Milkovich & Newman, 2007), meaning the remaining jobs contain configurations of task, duties, responsibilities, and behavioral requirements unique to each firm. I say this to underscore Johnson et al.'s observation

Correspondence concerning this article should be addressed to Craig J. Russell.
E-mail: cruss@ou.edu
Address: Price College of Business, University of Oklahoma, Norman, OK 73019

[1] One exception to this generalization is found historically in the assessment center literature. Target jobs are entry level and higher management positions, and 50% or more of ''applicants'' are current employees in individual contributor positions.

that an ideal synthetic validation database could transport validity to approximately 400% more jobs than addressed by ‘‘external’’ selection systems, although I suspect many of these jobs contain fewer employees and exhibit meaningful lower levels of job creation or turnover. Transportation of criterion validity via synthetic inferences to unique, firm-specific jobs will likely remain an unrealized opportunity until synthetic validity evidence is generally available.

Motor Oil — Yes, Time Motion Studies — No

Johnson et al. used an industrial engineering (IE) example, stating that synthetic validity was analogous to determining how long an entire task will take by summing the time-motion study estimates of time required for each of the task’s component parts. For this example to truly reflect how synthetic validity works, IEs would also have to survey incumbents in other jobs about task component frequency then estimate the time required for task completion in *those* jobs from measures of task component completion time measures obtained in the original job. This might be a safe IE inference if you can assume similar applicant pool KSA profiles, equipment, work environments (e.g., frequency of interruptions), and so forth. Having performed a number of time–motion studies early in my career, I cannot imagine an IE making those assumptions. Analogies and metaphors can help convey complex topics, and I wish this one did. It breaks down most severely when one considers the broad versus narrow construct domain issues discussed later in the article. The most compelling metaphor (at least, for me) revolves around the word ‘‘synthetic,’’ which I equate with ‘‘man-made’’ as opposed to based on local validation study evidence.

Mean-Based Synthetic Validity

Predictor mean-based synthetic validity inferences can be severely undermined by real differences in depth of labor market skill pools. Specifically, I have observed severe differences in test score means and standard deviations across geographically separated employment venues when working with national and international employers. Using the mean-based job component validity approach for a cohort of jobs common to each locale would have resulted in severely different test batteries and, if common tests had been suggested, severely different cut scores. Although ‘‘Test X’’ was forecast to exhibit criterion validity using the validity-based job component approach (and in fact did in one instance with local follow-up validity studies with $N > 10,000$), mean-based synthetic validity indicated that incumbents exhibited low, medium, and high ‘‘Test X’’ means across locales (i.e., significantly different ‘‘Test X’’ means). If ‘‘Test X’’ data had only been available from low-score locales, ‘‘Test X’’ would not have been included in the final test battery. Mean-based synthetic validity might be useful in placement decisions when mean differences in ‘‘Test X’’ are observed across jobs with different task demands, although this again assumes that applicants are drawn from a common labor pool when filling positions in diverse geographic locales.

Transporting Validity and the Uniform Guidelines

The *Uniform Guidelines on Employee Selection Procedures* (1978) Section 7B explicitly address evidence needed to ‘‘transport’’ criterion validity evidence obtained from another job to a target job. Specifically, it says:

> Criterion-related validity studies conducted by one test user, or described in test manuals and the professional literature, will be considered acceptable for use by another user when the following requirements are met: (1) Validity evidence. Evidence from the available studies meeting the standards of section 14B of this part clearly demonstrates that the selection procedure is valid; (2) Job

> similarity. The incumbents in the user's job and the incumbents in the job or group of jobs on which the validity study was conducted perform *substantially the same major work behaviors, as shown by appropriate job analyses* both on the job or group of jobs on which the validity study was performed and on the job for which the selection procedure is to be used; and (3) Fairness evidence. The studies include a study of test fairness for each race, sex, and ethnic group which constitutes a significant factor in the borrowing user's relevant labor market for the job or jobs in question. (*Uniform Guidelines for Employee Selection Procedures*, 1978, p. 206, emphasis added)

My concern has always been with the operational definition of ''substantially the same.'' Specifically, I envision a plaintiff's attorney asking the author of a synthetic validity study ''so, how many primary research studies in your 'validity-based' synthetic validity study actually looked like the job in question? Exactly how many studies had this profile of task importance ratings, this profile of behavioral importance ratings, this profile of task frequency ratings, this profile of behavioral frequency ratings? You didn't really do what was required in the *Guidelines* did you? You didn't take criterion validity evidence from a job that was ''substantially the same.'' You took validity evidence from a wide variety of jobs with very different job requirements and found a relationship between those job requirements and criterion validity. So really you just *estimated* what you expected criterion validity *might* be for this job from all these other, *very different* jobs.'' Rolling this imaginary dialog around in my head has caused me not to transport criterion validity evidence from Job A to Job B unless I have meaningful evidence suggesting substantial overlap in job requirements. Unfortunately, the *Guidelines* do not provide an operational definition of ''substantially the same.'' Evidence I have seen pass review by EEOC and OFCCP auditors (which does not mean it will ''pass review'' in the next audit), reflects subject-matter-expert (SME) agreement levels, precision of SME estimates, and similarity of job analysis questionnaire (JAQ) dimensional profiles. Specifically,

1. SME within group agreement on ratings of JAQ dimensional requirements has to meet heuristic minimum r_{wg} levels described in the literature (i.e., typically $r_{wg} > .43$; Kozlowski & Hattrup, 1992).
2. Adequate sample size required to assure 95% confidence intervals around any mean JAQ dimensional rating is less than 1.0. This means that no 95% confidence intervals (CI) derived for JAQ dimensional importance ratings includes more than one integer scale point on whatever ratings scale is being used—if 1.0 is ''somewhat important'' and 2.0 is ''Important,'' a sample size large enough to ensure that the 95% CI for a scale score with $\overline{X} = 2.0$ will not include 1.0 is required.
3. Jobs ''A'' and ''B'' must have the same profile of average job dimensional ratings exhibiting $\overline{X} = 2.0$, where the anchor for ''2'' ratings was ''important.'' In other words, the jobs have the same profile of job dimensions rated as ''important'' or ''extremely important'' (or whatever anchors were used for ratings higher than ''important''). Although Jobs A and B may vary in their profiles of what is ''important'' versus ''very important,'' criterion validity is not transported when Job A contains one or more key (i.e., ''important'' or higher) dimensions that are not ''important'' for Job B. Requiring identical profiles of JAQ dimensions rated ''important'' or higher essentially puts a floor on the severity of contamination inference errors about job content.

Note, only one of these three rules of thumb (and that is all they are) involves a test of statistical significance (e.g., H_0: $r_{\text{wg}} \leq .40$), the other two simply involve a judgment call by an investigator about how ''similar'' is similar enough to be called ''substantially the same.'' Note also that the three rules are very far removed from what might constitute the operational definition of ''exactly the same''—measurement equivalence of SME responses to JAQs in Jobs A and B. Importantly, it is also not the validity-based model of synthetic validity described by Johnson et al.

Conclusion

Again, I strongly agree with Johnson et al.'s main contentions. My points of disagreement revolve around the fact that (a) many more synthetic validity studies exist than what Johnson et al. suggest (although most are proprietary), (b) the IE metaphor does not serve the synthetic validity argument well, and (c) the *Uniform Guidelines* speak to synthetic validity in a more targeted way than Johnson et al. suggest. Johnson et al. made a number of other points that my comments also relate to, for example, discriminant validity is often evidenced in proprietary synthetic validity databases I have seen when the population of criterion validity studies wandered widely outside the cognitive predictor domain to include personality dimensions, assessment center ratings, and biographical information. I will leave it to the reader to make those connections. All in all, I share Johnson et al.'s excitement about synthetic validity's potential contributions to practice and theory and look forward to seeing even more reported in the publicly available literature.

References

Hunt, S. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology*, *49*, 51–84.

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.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, *77*, 161–167.

Milkovich, G., & Newman, J. (2007). *Compensation* (9th ed.). New York: McGraw Hill.

Uniform Guidelines on Employment Selection Procedures. (1978). *Federal Register*, *43*, 38290–38309.