Guidance for the Interpretation of Validity Coefficients

New research demonstrates issues with how validity coefficients are estimated.

Contents

Executive Summary

Authors

**Jeff Johnson,
Principal Scientist**

**Paul DeKoekkoek,
Science Director**

In talent assessment, the strongest form of validation evidence is generally considered to come from criterion-related validation, which demonstrates that scores on an assessment (i.e., the predictor) are related to scores on a criterion measure of interest (most often job performance). Criterion-related validity evidence is usually presented in the form of a *validity coefficient*, or correlation (*r*), which ranges from 0 to 1 and indicates the magnitude of the relationship between assessment and criterion scores.

There are two primary types of validation studies used to collect criterion-related evidence. *Concurrent validation* studies collect predictor data (i.e., assessment scores) from job incumbents and criterion data (e.g., manager ratings) on those incumbents close together in time. *Predictive validation* studies collect predictor data from job applicants prior to selection and criterion data from those hired after they have been on the job for some time. Each approach results in a validity coefficient that estimates the relationship between predictor and criterion scores. The accuracy of that validity coefficient depends on factors that differentially affect each criterion-related validation study, such as the reliability of the criterion measure. For example, performance ratings made by one rater will not be the same as ratings made by another rater because they have different perspectives. More reliable criterion measures tend to yield higher validity coefficients.

Range restriction can also suppress validity coefficients. This occurs when selection is based on assessment scores, because the range of scores on the criterion will be restricted due to criterion scores for those not selected not being available. This is direct range restriction on the predictor, which may be seen in a predictive validity study. In concurrent validation studies, there will be indirect range restrictions on both the predictor and the criterion because the incumbent group will have been selected on multiple variables that are related to the new predictor of interest, often to an unknown extent. Standard statistical formulas are available to correct validity coefficients for criterion unreliability, direct range restriction and indirect range restriction and these corrected coefficients provide a more precise estimate of the true relationship between assessment scores and criterion scores.

Meta-analysis is a method of cumulating results across multiple studies to get a more accurate estimate of the true correlation between variables than is possible in a single study. Corrections are made to validity coefficients from a meta-analysis, but they must often be made using estimates from other studies because the necessary information is not reported within each study. A recent article highlighted some common practices when making corrections in meta-analyses that can lead to overestimates of validity coefficients. The primary issues are that (a) concurrent validation studies make up the majority of studies in meta-analyses, but all validity coefficients are often corrected as though they come from predictive validity studies; and (b) reliability estimates differ across meta-analyses so the degree of correction for criterion unreliability is not consistent.

The general conclusion when comparing validity estimates from previous meta-analyses to validity estimates generated by more realistic and consistent estimates of the amount of range restriction and criterion reliability is that the validity of most selection procedures is not as high as previously believed. Nevertheless, the level of validity is still of practical use, and we must adjust our expectations for validity magnitude and look at validity claims with a critical eye. When adding potential adverse impact to the evaluation of talent assessment value, assessments like structured interviews and empirically keyed biodata measures are now seen as much more valuable than assessments of cognitive ability and work sample tests.

This report provides guidance for consumers of talent assessments as they evaluate claims of validity made by vendors. Extensively researched meta-analyses published in peer-reviewed academic journals often overestimate validity coefficients, making it likely that talent assessment vendors may make similar errors in their own research supporting the predictive ability of their assessments. We outline the different types of information that should be available from a vendor in a technical manual or other documentation and should be explained in a manner that is adequate for evaluating the appropriateness of the research and the resulting validity claims.

How Validity is Estimated

The term ***validity*** refers to the accuracy of interpretations drawn from assessment scores.[1,2] For example, common interpretations drawn from assessment scores include (a) the assessment content is job-related, (b) the assessment scores predict job performance, and (c) the assessment measures what it is intended to measure. ***Validation*** is the process of establishing evidence that supports the interpretation of assessment scores.

There are a variety of strategies that can be used to establish validation evidence. For example, content-oriented validation evidence is demonstrated when the content of an assessment representatively samples the important work behaviors, activities, and/or competencies necessary for job performance. Content validation relies heavily on expert judgment rather than statistical methods. These judgments are provided by subject matter experts (SMEs), who link competencies required on the job to competencies measured by an assessment.

In talent assessment, the strongest form of validation evidence is generally considered to come from criterion-related validation, which demonstrates that scores on an assessment (i.e., the predictor) are related to scores on a criterion measure of interest (most often job performance). Criterion-related validity evidence is usually presented in the form of a ***validity coefficient,*** or correlation (*r*), which ranges from 0 to 1 and indicates the magnitude of the relationship between assessment and criterion scores.

There are two primary types of validation studies used to collect criterion-related evidence. ***Concurrent validation*** studies collect predictor data (i.e., assessment scores) from job incumbents and criterion data (e.g., manager ratings) on those incumbents close together in time. ***Predictive validation*** studies collect predictor data from job applicants prior to selection and criterion data from those hired after they have been on the job for a period of time. Because of practical constraints and the need to demonstrate validity evidence prior to the operational use of an assessment, concurrent studies are more common than predictive studies. Each approach results in a validity coefficient that estimates the relationship between the predictor and criterion scores. The accuracy of that validity coefficient depends on many factors that differentially affect each criterion-related validation study.

[1] Society for Industrial and Organizational Psychology. "Principles for the validation and use of personnel selection procedures," *Industrial and Organizational Psychology* 11 (2018): 1-97.

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for educational and psychological testing* (Washington, DC: American Educational Research Association, 2014).

Factors That Affect the Validity Coefficient

The validity coefficient observed in any kind of criterion-related validation study is only an estimate of the true relationship between the predictor and the criterion. A variety of factors can influence the extent to which the observed validity coefficient is an accurate estimate of the true validity coefficient, among which are the following:

- **Sampling Error** – If the sample size is small and/or does not adequately represent the target population, the observed validity coefficient will be compromised by sampling error and is unlikely to accurately reflect the true relationship between variables. Sampling error can be minimized by ensuring representativeness and increase the sample size to the extent possible.
- **Moderator Variables** – In some cases, the strength of the relationship between two variables depends on a third variable, such as another personal characteristic or the work context. For example, a measure of agreeableness may be positively related to performance in a highly collaborative, team-based organizational culture, but may be negatively related to performance in an organization with a culture that is highly competitive and adversarial. The influence of moderator variables can be difficult to detect outside of a large study, so assessment users are advised to consider potential moderators and ensure that the assessment is appropriate for the situation.
- **Criterion Unreliability** – Neither assessment nor criterion scores can be perfect reflections of an individual's true scores on those measures because all psychological assessments are measured with some degree of unreliability. If an individual were to complete an assessment multiple times with no learning in between, the scores would not be exactly the same each time due to factors such as differences in mood, amount of sleep, random errors, or unstable aspects of the assessment itself. Similarly, a performance rating instrument completed by one rater will reflect only the behaviors observed by the rater. Another rater with a different perspective would likely provide somewhat different ratings. Neither rater can observe all possible behaviors that are relevant to the criterion measure.

The validity coefficient is suppressed to the extent that variables are not perfectly reliable. Therefore, we do all we can to make our assessments and criterion measures as reliable as possible. For performance ratings, we can enhance reliability by designing the rating scale to be interpreted as objectively as possible, providing rater training to give each rater the same frame of reference, collecting ratings for research purposes only, and using multiple raters whenever possible. In addition, a simple correction formula is available that allows researchers to estimate the size of the validity coefficient if the criterion measure were perfectly reliable:

$$r_{xyT} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

where r_{xyT} is the true validity estimate, r_{xy} is the observed validity coefficient and r_{yy} is the reliability of the criterion measure.

Note that in selection settings, we would not correct for unreliability in the predictor because selection must come from the actual assessment scores, not a theoretical perfectly reliable assessment.

Range Restriction – To get the best estimate of a true correlation between a predictor and a criterion, there should be a wide range of scores from the low end to the high end on both variables. This is rarely possible in real-world talent assessment research because there will usually be some degree of selection on the predictor and/or the criterion that restricts the range of observed scores. If selection is based entirely on assessment scores, the range of scores on the criterion will be restricted because the criterion scores for those not selected are not available. This is direct range restriction on the predictor, which may be seen in a predictive validity study. Given the ratio of the standard deviation in the selected group to the standard deviation in the unselected group (the U ratio), a correction formula can be applied to adjust the observed correlation in the selected group to obtain a better estimate of the correlation in the population.

In concurrent validation studies, there will be indirect range restriction on both the predictor and the criterion because the incumbent group will have been selected on multiple variables that are related to the new predictor of interest, often to an unknown extent. Those who are currently on the job may have completed one or more assessments, with selection based on some weighted composite or a multiple hurdle approach. Criterion variance may be restricted as experience reduces performance differences among incumbents. Also, high performers may be promoted out of the job while low performers may have been selected out. A different correction formula is available that takes into account both the U ratio and the correlation between the predictor and the selection variable used to select incumbents.

SHL's Approach

SHL has historically taken a conservative approach to correct validity coefficients. We always report uncorrected validity coefficients and explain the procedure when any corrections are made. Corrections for criterion unreliability are based on interrater reliability estimates for performance ratings if multiple raters are available for a reasonable sample of cases. When all ratings are made by a single rater, we will often use estimates from meta-analyses of performance rating reliability,[3,4] using a more conservative value rather than the typical value. Corrections for range restriction are typically not conducted, primarily because most criterion-related validation studies are concurrent and, therefore, realistic estimates of correction formula variables are difficult to make. For predictive studies in which all relevant information is available, we do report validity coefficients that are corrected for range restriction with a complete explanation for how the corrections were made.

[3] J. M. Conway and A. I. Huffcutt. "Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings," *Human Performance* 10 (1997): 331-360.
[4] C. Viswesvaran, D. S. Ones, and F. L. Schmidt. "Comparative analysis of the reliability of job performance ratings," *Journal of Applied Psychology* 81 (1996):557-574.

Meta-analysis

Meta-analysis is a method of cumulating results across multiple studies to reduce the influence of sampling error and get a more accurate estimate of the true correlation between variables than is possible in a single study. Because different studies will differ in the amount of range restriction or criterion unreliability, corrections are made to each study to better equate estimates across studies. Meta-analyses are ubiquitous in the academic literature for a wide variety of constructs or methodologies used for selection. For example, one meta-analysis[5] computed the mean validity coefficient across 258 studies examining the relationship between overall assessment center ratings and leader overall performance. Large talent assessment companies that have conducted many criterion-related validation studies for a particular assessment will often conduct a meta-analysis to get a more stable estimate of the validity coefficient.

Schmidt and Hunter (1998)[6] summarized meta-analyses of selection procedures that had been conducted to that time. This article has been considered the last word on the level of validity for different selection procedures (as of mid-2022, it has been cited over 6,400 times). Many of the meta-analyses summarized in this article are quite old, however, and many updated meta-analyses have been conducted since that time that present a more accurate picture. For example, an updated meta-analysis of work sample test validity[7] reported a mean corrected validity coefficient of .33, compared to the .54 reported in an earlier meta-analysis[8] and summarized in Schmidt and Hunter (1998). Nevertheless, in the nearly 25 years since this article was first published, many assessment consumers have come to expect the same type of validity for different types of assessments that are seen in Schmidt and Hunter (1998).

[5] W. Arthur, E. A. Day, T. L. McNelly, and P. S. Edens. "A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology* 56 (2003): 125-154.

[6] F. L. Schmidt and J. E. Hunter. "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings," *Journal of Applied Psychology* 124 (1998): 262–274.

[7] P. L. Roth, P. Bobko, and L. A. McFarland. "A meta-analysis of work sample test validity: Updating and integrating some classic literature," *Personnel Psychology* 58 (2005): 1009-1037.

[8] J. E. Hunter and R. F. Hunter. "Validity and utility of alternative predictors of job performance," *Psychological Bulletin* 96 (1984): 72-98.

Problems with Corrections in Meta-analysis

Given this context, a recent article by Sackett et al. (2022)[9] highlighted some common practices when making corrections in meta-analyses that can lead to overestimates of validity coefficients. By extension, this article calls into question the perhaps overly optimistic values presented in Schmidt and Hunter (1998). These practices are primarily associated with estimating artifact distributions when the necessary information for computing corrections is not available for each study included in the meta-analysis. Ideally, corrections for range restriction and criterion unreliability would be made within each study and the mean of these corrected validity coefficients would then be computed across studies (weighted by sample size). Because most studies do not include the information needed for correcting the validity coefficients (e.g., interrater reliability, variance in the unselected group), an artifact distribution is typically created from the subset of studies that do include the relevant information. Based on the assumption that the amount of range restriction observed in this subset of studies is representative of the entire set of observed validities, the mean and variance of the entire set are corrected using this artifact distribution.

Range Restriction

Sackett et al. (2022) pointed out that most meta-analyses contain both predictive and concurrent studies. The U ratio (selected group SD/unselected group SD) may be known in predictive studies but is rarely known in concurrent studies. The smaller the U ratio, the more range restriction is present. Surprisingly, it is common practice to apply the mean U ratio obtained from primarily predictive studies to the mean observed validity coefficient across studies that are primarily concurrent. Sackett et al. demonstrated that the U ratio for concurrent studies is likely to be close to 1.0, indicating almost no range restriction, because correlations between the predictor variable of interest and the unknown selection variable are almost always very small. The effect of indirect range restriction in concurrent studies is therefore likely to be negligible. (For the same reasons, range restriction on the criterion variable is also likely to be minimal.) The effect of applying a U ratio based on predictive studies to the correction formula for indirect range restriction in concurrent studies is therefore an overcorrection. Given that most validity coefficients in a meta-analysis are usually based on concurrent studies, meta-analytic estimates of validity coefficients will likely overestimate the true validity coefficient.

Sackett et al. (2022) evaluated four other methods of estimating a U ratio and found those methods to also lead to overestimates of true validity coefficients due to untenable assumptions (e.g., unrealistically low selection ratios, and selection based on a single variable).

Criterion Unreliability

Corrections for criterion unreliability also tend to be inconsistent across meta-analyses. Reliability estimates that come from a representative sample of studies included in the meta-analysis are reasonable, but many meta-analysts are forced to estimate reliability based on distributions borrowed from other settings. Several meta-analyses have reported a mean interrater reliability of .52 for supervisor ratings of overall performance.[10,11,12,13] Many meta-analyses have used this value to make criterion unreliability corrections, but many others have used .60. If different meta-analyses base validity estimates on different assumptions of

[9] P. R. Sackett, C. Zhang, C. M. Berry, and F. Lievens. "Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range." *Journal of Applied Psychology* 107 (2022): 2040-2068. https://doi.org/10.1037/apl0000994

[10] Conway and Huffcutt (1997)

[11] J. F. Salgado, N. Anderson, S. Moscoso, C. Bertua, F. De Fruyt, and J. P. Rolland. "A meta-analytic study of general mental ability validity for different occupations in the *European community," Journal of Applied Psychology* 88 (2003): 1068–1081.

[12] J. F. Salgado and G. Tauriz. "The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies," *European Journal of Work and Organizational Psychology* 23 (2014): 3–30.

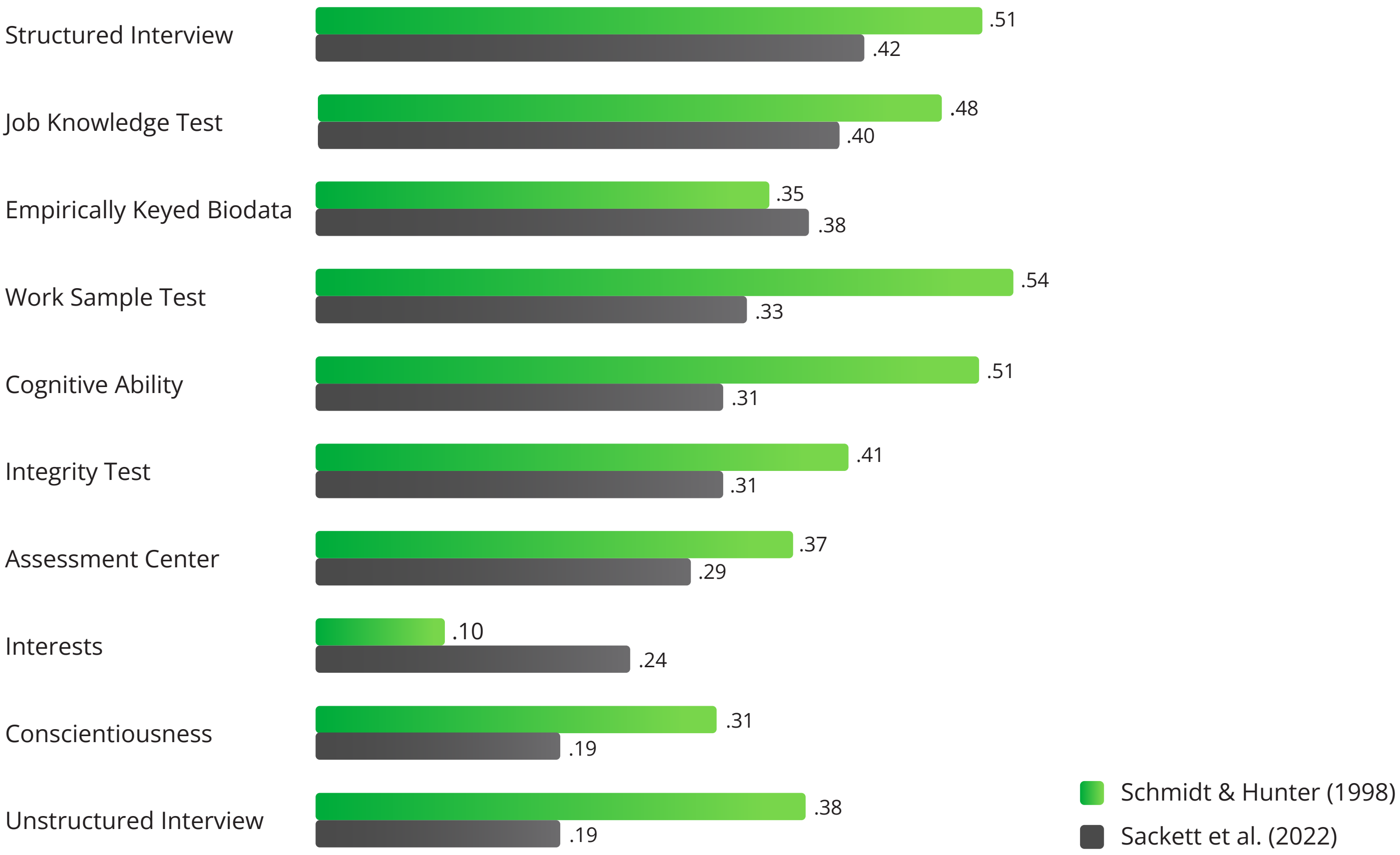
[13] Viswesvaran et al. (1996)

criterion reliability, it is difficult to compare the relative validity of the same or similar predictors, let alone different predictors. The .60 value is based more on assumptions than on data, but there are reasons to select .60. This was the value most often used in early meta-analyses based on assumed distributions.[14,15] One meta-analysis found interrater reliability across jobs ranging from .48 to .60 depending on job complexity.[16] Another study found that interrater reliability steadily increased with increased opportunity to observe, up to an asymptote of about .60.[17] Sackett et al. (2022) proposed using a consistent value of .60, which is the most conservative reasonable estimate.

Implications

Because of these results, Sackett et al. (2022) re-examined the meta-analyses summarized by Schmidt and Hunter (1998) and other more recent meta-analyses to generate new estimates of meta-analytic validity coefficients based on revised range restriction corrections (or no correction if determined to be most appropriate) and consistent estimates of criterion unreliability (results for criterion reliability of both .52 and .60 were included). The original and revised validity estimates for selected assessment types are shown in the figure below, with revised estimates based on criterion reliability of .60:

Original and Revised Validity Estimates for Assessment Types



[14] Hunter and Hunter (1984)

[15] K. Pearlman, F. L. Schmidt, and J. E. Hunter. "Validity generalization results for tests used to predict job profi-ciency and training success in clerical occupations," *Journal of Applied Psychology* 65 (1980): 373–406.

[16] Conway and Huffcutt (1997)

[17] H. R. Rothstein. "Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe," *Journal of Applied Psychology* 75 (1990): 322-327.

The general conclusion when comparing validity estimates summarized by Schmidt and Hunter (1998) to validity estimates generated by Sackett et al. (2022) is that the validity of most selection procedures is not as high as previously believed. Nevertheless, the level of validity is still of practical use, and we just need to adjust our expectations for validity magnitude and look at validity claims with a critical eye.

When comparing the relative magnitude of validity estimates for different assessment types, we can conclude that what were considered the most valid selection procedures previously are still generally the most valid following the re-analysis, but the rank order changes. Cognitive ability tests and work samples decreased in estimated validity the most (each dropping by at least .20), while structured interviews would now be considered the most valid assessment type. An important result is that the difference in validity between structured and unstructured interviews is now much bigger than was indicated by Schmidt and Hunter (1998). Their review had validity estimates of .51 for structured interviews and .38 for unstructured interviews. Sackett et al.'s (2022) reanalysis yields validity estimates of .42 and .19 for structured and unstructured interviews, respectively.

Some validity estimates did increase following the reanalysis, based on more recent data and more appropriate assumptions about what studies should be included and how they should be treated. For example, biodata assessments were split into empirically keyed and rationally keyed versions based on a meta-analysis that found the scoring method was a moderator variable,[18] demonstrating that empirically keyed biodata has higher validity (.40) than the .35 that was reported in an earlier meta-analysis[19] across all biodata assessments. As another example, Sackett et al. (2022) reported a higher validity coefficient for interest inventories (.24 vs. .10) because they examined the relationship between interests and job performance in jobs for which the interest profile was relevant rather than examining this relationship across all jobs.

[18] A. Speer, C. Sendra, and M. Shihadeh. *Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates.* Paper presented at the 36th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA, April, 2021.

[19] H. R. Rothstein, F. L. Schmidt, F. W. Erwin, W. A. Owens, and C. P. Sparks. "Biographical data in employment selection: Can validities be made generalizable?" *Journal of Applied Psychology* 75 (1990): 175–184.

Variance of Validity Estimates

With meta-analyses of validity coefficients, the focus is typically on the mean (i.e., what is the average level of validity expected from an assessment?). Equally important, however, is the variance in the validity coefficients (i.e., how much variability is there across studies?). A meta-analysis may show that the mean of validity coefficients for an assessment may be very low, but if the variance is high, that may indicate that the assessment has strong validity in certain situations. The ***credibility interval*** is computed from the standard deviation of corrected estimates across studies and indicates the extent to which moderators may influence the level of validity. The credibility interval is not to be confused with the ***confidence interval,*** which measures the extent to which sampling error influences the accuracy of the effect.[20] When there is a large credibility interval, it is necessary to identify what characteristics of different studies may account for differences in validity. For example, the congruence of interests with the requirements of the job moderates the relationship between interests and job performance.[21]

Meta-analyses of personality assessment validity indicate that substantive moderators affect the magnitude of the validity coefficient. In other words, situational specificity seems to be the rule for the validity of personality assessments, as opposed to the highly generalizable validities seen with most other talent assessments. A summary of meta-analysis results for Big Five validities found an average 80% credibility interval of 0.30.[22] If there were no moderators involved, the credibility interval would be near zero. This large average credibility interval indicates that a great deal of variability in estimated validity across studies is not accounted for by sampling or measurement error. In fact, personality scales that are positively related to a performance dimension in some situations may have legitimately negative correlations with the same performance dimension in other situations. For example, a person high in Agreeableness may do well in an organization that has a team-based, cooperative culture but may have difficulty in an organization with a culture that is highly competitive and adversarial.

Leadership Validation Study

SHL has done extensive research on the moderators of personality assessment validity.[23] Between 2014 and 2016, SHL conducted the largest validation study of its type to define a taxonomy of organizational context factors and investigate its impact on predicting leader performance from personality. The Leadership Validation Study (LVS) included nearly 8,700 leaders, 5,900 supervisors, and over 33,000 direct reports from 85 companies representing more than 25 industries globally. Leader personality was measured with the Occupational Personality Questionnaire (OPQ) and performance was measured with a multisource performance rating instrument completed by each leader's supervisor and direct reports.

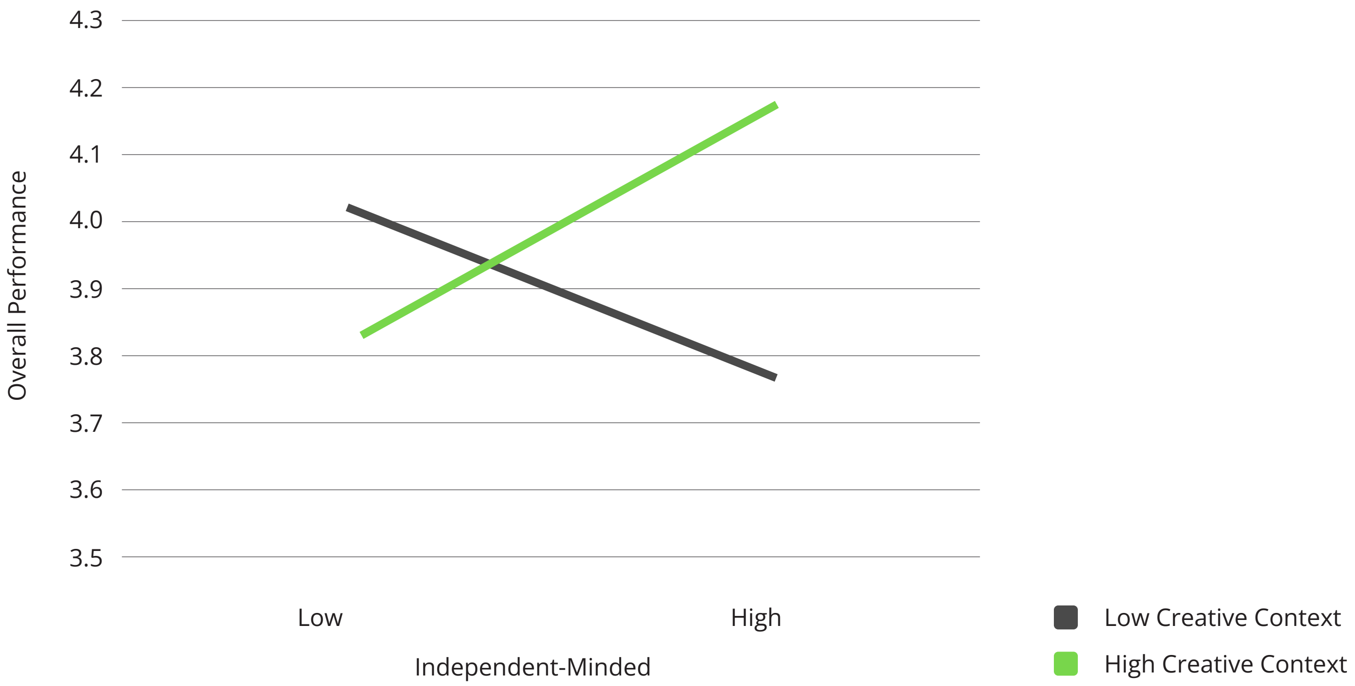
All participants completed surveys that were used to define the leader's work context. For example, leaders completed a survey to identify the most important aspects of their unique roles. Supervisors completed a survey measuring business priorities and different aspects of the organizational culture. Direct reports completed a survey measuring team functioning and characteristics. We created numerous context variables from these data that describe the unique work environment for any particular leader at the role, team, and organization level. Role-level contexts include aspects of the leader's job that often differ from role to role (e.g., the extent to which designing and driving new strategies is important to the job). Team-level contexts include the dynamics and makeup of the team, such

[20] E. M. Whitener. "Confusion of confidence intervals and credibility intervals in meta-analysis," *Journal of Applied Psychology* 75 (1990): 315-321.
[21] C. D. Nye, R. Su, J. Rounds, and F. Drasgow. "Interest congruence and performance: Revisiting recent meta-analytic findings," *Journal of Vocational Interests* 98 (2017): 138-151
[22] R. P. Tett and N. D. Christiansen. "Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007)," *Personnel Psychology* 60 (2007): 967-993.
[23] *SHL Leader Edge technical manual*. (Washington, DC: Author, 2018).

as the need to transform a team with a high-conflict culture. Organization-level contexts include the business priorities and culture of the organization (e.g., the extent to which growing the business through innovation is a priority).

We found that taking context into account brings increased precision in measurement and prediction. Leader success is greatly influenced by context, in that the personality traits that predict performance depend on the context in which the leader works. Predicting leader performance within contexts gave us three times better prediction on average than was possible when we did not incorporate context. For example, we found that being independent-minded predicts leader performance in opposite directions depending on the level of importance placed on creating an environment that consistently yields creative and innovative ideas, products, or services from team members. When driving creativity is important, more independent-minded leaders tend to be seen as better performers.
When driving creativity is less important, being independent-minded is less valued and going along with the crowd tends to lead to perceptions of better performance.

Example of the Impact of Context on Prediction



Predicting leader performance within contexts gave us three times better prediction on average than was possible when we did not incorporate context.

Taking multiple contexts into account improved prediction even further. Across 40 iterations of nine randomly selected contexts, we found a mean validity coefficient of .43 (corrected for criterion unreliability; .31 uncorrected). This value exceeds the validity found in meta-analyses of other commonly used leader assessments such as general cognitive ability (.25, corrected for criterion unreliability and range restriction),[24, 25] situational judgment tests (.28, corrected for criterion unreliability),[26] and assessment centers (.36, corrected for criterion unreliability and range restriction).[27]

Adverse Impact

When evaluating validity coefficients, the level of validity must be balanced against potential adverse impact when evaluating the value of an assessment. An assessment that produces high levels of prediction may not be very valuable if it also produces high levels of adverse impact. In the past, the validity of cognitive ability tests was often considered to be so much better than other potential predictors that adverse impact was considered worth the tradeoff. As research on talent assessment has matured, assessments with comparable validity and less adverse impact have been identified.

When examining Black-White standardized mean differences (d) along with updated validity coefficients, structured interviews, empirically keyed biodata, integrity tests, and personality assessments have comparable or greater validity than cognitive ability tests and much less potential adverse impact.[28]

[24] T. A. Judge, A. E. Colbert, and R. Ilies. “Intelligence and leadership: A quantitative review and test of theoretical propositions,” *Journal of Applied Psychology* 89 (2004): 542-552.

[25] B. J. Hoffman, D. J. Woehr, R. Maldagen-Youngjohn, and B. D. Lyons. “Great man or great myth? A quantitative review of the relationship between individual differences and leader effectiveness,” *Journal of Occupational and Organizational Psychology* 84 (2011): 347-381.

[26] M. S. Christian, B. D. Edwards, and J. C. Bradley. “Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities,” *Personnel Psychology* 63 (2010): 83-117.

[27] W. Arthur, E. A. Day, T. L. McNelly, and P. S. Edens. “A meta-analysis of the criterion-related validity of assessment center dimensions,” *Personnel Psychology* 56 (2003): 125-154.

[28] Sackett et al. (2021)

Guidelines for Evaluating Validity Coefficients

The purpose of this section is to provide guidance for consumers of talent assessments as they evaluate claims of validity made by vendors. We have seen that extensively researched meta-analyses published in peer-reviewed academic journals have often overestimated validity coefficients through the misapplication of artifact correction formulas.[29] This makes it likely that talent assessment vendors may make similar errors in their own research supporting the predictive ability of their assessments. The information outlined in this section should be available from a vendor in a technical manual or other documentation and should be explained in a manner that is adequate for evaluating the appropriateness of the research and the resulting validity claims.

Type of Study

The research reported could be based on a single study or multiple studies. A single study may be conducted in one organization or using the same measures across multiple organizations (i.e., a ***consortium study***). A consortium study may be conducted if adequate data are not available within a single organization. SHL's LVS study demonstrating the impact of context on personality assessment validity is an example of a consortium study.

If multiple studies are used to support the validity of an assessment, each study may be reported separately, or the results of each study may be combined quantitatively through a meta-analysis. The similarities and differences across multiple studies should be described (e.g., type of organization, occupation, criterion measure, predictive or concurrent study). When corrections for range restriction are applied, it is especially important to understand the mix of predictive and concurrent studies.

Sample

The validation study sample should be representative of the population to which the assessment is applied operationally. For example, an assessment developed for use in selecting customer-facing retail workers would ideally not be validated on a sample of office workers, unless the assessment is designed to predict behavior that is common across the two professions. If a validation study conducted in one setting is used to support the use of an assessment in another setting, the U.S. Equal Employment Opportunity Commission (EEOC) requires that the jobs in each setting share substantially the same major work behaviors.[30]

The sample size should be adequate for identifying a relatively stable validity coefficient. The sample should be large enough that the validity coefficient is statistically significant, but statistical significance is not sufficient for concluding that the validity coefficient is meaningful. A small validity coefficient (e.g., < .10) is more and more likely to achieve statistical significance as the sample size increases, so statistical significance is a minimum but not sufficient requirement for concluding that a single validity coefficient is meaningful.

To assess the adequacy of the sample size associated with a validity coefficient, a confidence interval should be computed around the point estimate. If a confidence interval is not provided in the technical documentation, it can be computed using an online calculator or very closely approximated by computing a simple approximation of the standard error of the correlation coefficient (SE_r):

$$SE_r = \frac{1 - r^2}{\sqrt{N - 2}}$$

where r is the validity coefficient and N is the sample size. A 95% confidence interval can be computed by

[29] Ibid

[30] Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. "Uniform guidelines on employee selection procedures," *Federal Register* 43 (1978): 38294-38309.

multiplying SE_r by 2 and subtracting the product from r to get the lower end of the confidence interval and adding the product to r to get the higher end of the confidence interval. For example, if a validity coefficient of .20 is calculated from a sample of 100, the standard error would be:

$$SE_r = \frac{1 - .20^2}{\sqrt{100 - 2}} = \frac{.96}{9.9} = .097$$

The 95% confidence interval would be $r \pm 2 \times SE_r$, so $.006 < r < .394$. The confidence interval is interpreted as the range within which we can be 95% confident that the true population correlation lies. In this example, the true value of the validity coefficient is very likely anywhere between .01 and .39, which is a very wide range that indicates the assessment may be one of the most predictive assessments at best or may have virtually no predictive value at worst. A 95% confidence interval that contains 0 indicates the correlation is not significant at $p < .05$ in a two-tailed significance test. Note that an 80% confidence interval is commonly computed for meta-analytic correlations, and this should be reported in the documentation of the meta-analysis.

To decrease the standard error and increase the confidence in the estimate of a validity coefficient, the sample size should be as large as possible while still being representative of the population of interest. Because large sample sizes are often difficult to procure in validation studies, researchers will use other methods to increase sample size, such as consortium studies, meta-analyses, or synthetic validation techniques.[31]

Criterion

The quality of the criterion variable (i.e., what we are trying to predict) is a critical but often overlooked aspect of a validation study. The criterion is most often obtained from ratings of job performance made by someone with the opportunity to observe relevant behavior but may also be a more objective measure such as sales metrics, productivity per hour, or turnover. Ratings tend to be preferred because of their flexibility, as a rating scale can be designed for any performance construct that can be described in behavioral terms. "Objective" measures may still involve some judgment, may be influenced by factors outside the individual's control, and are likely to measure only a small part of the performance domain for a job.[32] Rating scales can be written for all relevant aspects of the job and can evaluate the behavior of the individual rather than the effectiveness of an outcome.

Most validation studies include a measure of overall performance, which may be measured through one or more direct overall performance judgments or by adding ratings made on more specific performance dimensions. If the latter, performance dimensions should be demonstrated to be job-related through a job analysis or competency modeling exercise. The broadest performance dimensions beneath overall performance include task performance, citizenship performance, and adaptive performance. Task performance consists of the technical proficiency aspects that separate one job from another, such as developing assessments, writing code, or making decisions.[33] Citizenship performance is behavior that supports the broader organizational environment, including helping and cooperating with others, representing the organization favorably, and demonstrating effort and initiative.[34] Adaptive performance is behavioral change in response to an

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

[32] W. C. Borman, M. R. Grossman, R. H. Bryant, and J. Dorio. "The measurement of task performance as criteria in selection research," In J. L. Farr and N. T. Tippins (Eds.), *Handbook of Employee Selection* (2nd Ed., Ch. 20). (2017): 429-447.

[33] Borman et al. (2017).

[34] D. W. Dorsey, J. M. Cortina, M. T. Allen, S. D. Waters, J. P. Green, and J. Luchman. "Adaptive and citizenship-related behaviors at work," In J. L. Farr and N. T. Tippins (Eds.), *Handbook of Employee Selection* (2nd Ed., Ch. 21). (2017): 448-475.

altered situation, including recognizing situational demands and taking needed actions.[35] More specific dimensions are subsumed under these broader dimensions, such as SHL's 20-dimension Universal Competency Framework.[36]

An additional criterion construct that is somewhat unique from traditional performance dimensions is counterproductive work behavior, which is a negative aspect of performance. These are behaviors that run counter to the legitimate interests of an organization, such as stealing, damaging property, leaving early, or abusing coworkers.[37] Correlations between assessment scores and scores on each performance dimension measured should be reported, not just correlations with overall performance. The assessment may not necessarily be relevant to every performance dimension but examining dimension-level correlations helps to understand if the assessment is predicting what it is expected to predict.

Performance ratings are most often provided by the target person's immediate supervisor, but ratings may also be provided by peers and/or subordinates. Raters from these different perspectives have opportunities to observe different behaviors and ratings collected from multiple perspectives can provide a more complete picture of the individual's total performance.[38] Rater perspective also influences interrater reliability, as research shows that supervisors tend to provide the most reliable ratings, followed by peers, then subordinates.[39] It is much easier to obtain multiple ratings from the peer and subordinate perspectives, but, mean ratings collected from three or four peers or subordinates will be more reliable than a single rating from a supervisor.[40] It is important to understand who provided the ratings, how they were combined into a single score, and whether the proper estimate of interrater reliability was applied.

When the validity coefficient is based on a meta-analysis, the criterion variable must be consistent across studies. If some studies use a task performance criterion and others use a citizenship performance criterion, the meta-analytic correlation would be impossible to interpret. Even if the same criterion label is used in every study (e.g., overall performance), it may not make sense to include all studies. If different studies are based on different jobs (e.g., firefighter vs. retail worker), "overall performance" may mean something different in each study if the tasks performed and evaluated are not similar.

Finally, the purpose of the ratings should be made clear. Ratings that are made specifically for the purpose of a validation study are preferred over ratings that are made for administrative purposes (e.g., annual performance reviews). Administrative ratings are often influenced by factors other than the true performance of the individual (e.g., a desire to motivate a subordinate, maintaining interpersonal relationships, and justifying salary or promotion decisions).[41]

Corrections

Criterion Unreliability

When validity coefficients are corrected for criterion unreliability, technical documentation must make clear where the reliability estimate came from and how it was computed. If the validation study (or a subset of studies in a meta-analysis) has the appropriate data for computing interrater reliability, that is the value that should be used to make the correction. If an estimate is used based on another set of data (e.g., a published meta-analysis), a clear rationale for why that value was chosen should be provided.

[35] Ibid.

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

[37] M. Rotundo and P. E. Spector. "New perspectives on counterproductive work behavior including withdrawal," In J. L. Farr and N. T. Tippins (Eds.), *Handbook of Employee Selection* (2nd Ed., Ch. 22). (2017): 476-508.

[38] W. C. Borman. "The rating of individuals in organizations: An alternative approach," *Organizational Behavior and Human Performance* 12 (1974): 105-124.

[39] Conway and Huffcutt (1997)

[40] G. J. Greguras and C. Robie. "A new look at within-source interrater reliability of 360-degree feedback ratings," *Journal of Applied Psychology* 83 (1998): 960-968.

[41] F. J. Landy and J. L. Farr. "Performance rating," *Psychological Bulletin* 87 (1980): 72-107

There are a variety of different types of reliability, including internal consistency (e.g., coefficient alpha), stability over time (e.g., test-retest correlations), and interrater reliability. For performance ratings, we are interested in the question of the extent to which the same ratings would be obtained if ratings were obtained from a different rater, so interrater reliability is most appropriate.[42] There are several measures of interrater reliability or interrater agreement that may be applied. Below we describe alternative measures and when they are most appropriate:

- **Interrater correlation** – The Pearson correlation between ratings made by two raters, or the mean correlation between three or more raters, is the most commonly used measure of interrater reliability.[43] The correlation coefficient measures the extent to which different raters provide the same pattern of high and low ratings.

  This type of correlation is appropriate when two or more raters provide ratings on each of the same set of ratees. Although some have argued that the interrater correlation is not a measure of reliability,[44] others have demonstrated that this argument confuses reliability with construct validity [45] and that the correlation between the ratings made by two raters captures both errors of judgment and idiosyncratic rater perceptions.[46]

- **Intraclass correlation** – There are six different forms of the intraclass correlation, the appropriateness of each depending on whether (a) each target is rated by a different set of raters or all raters rate all targets, (b) raters are considered to be a random sample of a larger population of potential raters, and (c) the unit of analysis is an individual rating or the mean of multiple ratings.[47] Depending on each of these factors, different elements from an analysis of variance (ANOVA) are input into an appropriate equation.

  For performance ratings in a validation study, each target is typically rated by a different set of raters who represent a sample of a larger potential pool of raters, and we are interested in the mean rating. *ICC*(1, *k*) is the appropriate intraclass correlation in this situation. *ICC*(1, *k*) evaluates the ratio of variance within groups to variance between groups to measure the extent to which those rating one person have less variance in their ratings than is seen across all raters and all targets. It is computed as ($MS_{between}$ – MS_{within}) / $MS_{between}$, where MS is the mean square from a one-way ANOVA. If more than one rater rates the same set of targets (as with the Pearson correlation), *ICC*(2, *k*) provides a measure that takes into account the similarity in both the pattern and absolute level of the ratings made by two raters[48] This is a more complete measure of consistency across raters but is less often used than the simple Pearson correlation.

- ***r_{wg}*** – When multiple raters rate a single target, r_{wg} provides a measure of the extent to which raters agree compared to the level of agreement that would be expected by chance.[49] This index takes the expected distribution of ratings into account when determining the level of agreement. For example, performance ratings tend to be skewed toward the positive end of the scale, so a baseline level of agreement could be expected simply because of shared rating tendencies. Accounting for this with the null distribution yields a more accurate picture of the level of agreement. The maximum possible value for r_{wg} is 1.0, which results when all ratings are the same. There is no widely accepted standard for what constitutes acceptable interrater agreement but, typically, an $r_{wg} \geq .70$ is considered good agreement, an r_{wg} in the .40 to .60 range represents a moderate agreement, and an $r_{wg} \leq .40$ represents low agreement.[50] Note that interrater agreement differs from interrater reliability in that

[42] Viswesvaran et al. (1996)

[43] Sackett et al. (2022)

[44] K. R. Murphy and R. DeShon. "Interrater correlations do not estimate the reliability of job performance ratings," Personnel Psychology 53 (2000): 873-900.

[45] F. L. Schmidt, C. Viswesvaran, and D. S. Ones. "Reliability is not validity and validity is not reliability," *Personnel Psychology* 53 (2000): 901-912.

[46] J. E. Hunter and F. L. Schmidt. *Methods of meta-analysis* (2nd Ed.). (Thousand Oaks, CA: Sage, 2004).

[47] P. E. Shrout and J. L. Fleiss. "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin* 86 (1979): 420-428.

[48] Shrout and Fleiss (1979)

[49] L. R. James, R. G. Demaree, and G. Wolf, "Estimating within-group interrater reliability with and without response bias," *Journal of Applied Psychology* 69 (1984): 85-98.

agreement considers similarity in the absolute level of ratings whereas reliability focuses more on agreement in the pattern of ratings. Therefore, r_{wg} should not be used when correcting for criterion unreliability.

- **Generalizability coefficient** – Generalizability theory[51] allows the reliability coefficient to be deconstructed to assess the magnitude of multiple sources of variance. For example, ratings from multiple raters on a multiple-item performance rating instrument will have variance due to the items on the instrument, variance within individual raters, and variance across different raters. Generalizability coefficients take all these sources of variance into account to evaluate the extent to which scores contain measurement error. Generalizability coefficients are rarely reported in technical manuals or published research, but they provide more information than traditional reliability coefficients.

In sum, the most common and appropriate interrater reliability coefficients for criterion unreliability corrections are *ICC*(2, *k*) and the simple Pearson correlation when multiple raters rate the same set of ratees, and *ICC*(1, *k*) when each ratee is rated by a different set of raters. The generalizability coefficient is also appropriate but much more difficult to compute. The r_{wg} index is a useful measure of interrater agreement in rating a single target but should not be used for unreliability corrections.

One might argue that range restriction can affect correlations between criterion ratings just as it affects correlations between predictor scores and criterion scores,[52] as high performers get promoted and poor performers leave the job. Criterion range restriction would be indirect because selection out of a job would be based on factors only weakly related to the assessment scores on which a selection decision was made, so the argument for little indirect range restriction on the predictor holds for any possible range restriction on the criterion as well.[53]

Finally, consumers of talent assessments should apply the correction formula to the reported uncorrected correlation using the reported reliability to ensure that the correction was computed correctly.

Range Restriction

When evaluating a correction for range restriction, it is essential to determine whether the correct formula has been applied. The formula for direct range restriction should only be applied when selection is based only on scores on the assessment for which the validity coefficient is calculated (i.e., predictive studies). This is rarely the case in most validation studies, and direct range restriction corrections should never be applied indiscriminately with a meta-analysis that includes mostly concurrent studies. The formula for indirect range restriction may be applied with concurrent studies, but this presumes that the correlation between the assessment variable and the variable on which previous selection into the job was made is known. Sackett et al. (2022) showed that this correlation is unlikely to be large enough to have a meaningful influence on indirect range restriction corrections.

If claims are made about direct or indirect selection variables, it is appropriate to ask questions about other potential selection variables that may influence the extent of range restriction. For example, if a direct range restriction correction is applied because selection is claimed to be based on scores on a single assessment, one should inquire whether any other factors could have influenced the final selection decision, such as an interview or a resume screen.

[50] S. W. Kozlowski and K. Hattrup. "A disagreement about within-group agreement: Disentangling issues of consistency versus consensus," *Journal of Applied Psychology* 77 (1992): 161-167.

[51] L. J. Cronbach, G. C. Gleser, H. Nanda, and N. Rajaratnam. *The dependability of behavioural measurements: Theory of generalizability for scores and profiles* (New York: Wiley, 1972).

[52] P. R. Sackett, R. M. Laczo, and R. D. Arvey. "The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research," Personnel Psychology 55 (2002): 807-825.

[53] Sackett et al. (2022)

It is also important to determine where the unselected group standard deviation (SD) came from. Was this value computed directly from study data or borrowed from another study or studies? If a meta-analysis is reported, ensure that the same value was not applied to all studies regardless of type. The ratio of the selected group SD to the unselected group SD will decrease with the selection ratio (percentage of applicants accepted). Determine whether the SD ratio makes sense given the selection ratio, which should be reported.

Joint Corrections for Criterion Unreliability and Range Restriction

Many validity coefficients may be corrected for both criterion unreliability and range restriction. When direct range restriction has occurred, the appropriate order is to correct the validity coefficient for unreliability in the criterion first and then apply the range restriction correction to the criterion unreliability-corrected correlation.[54] In the case of indirect range restriction, the process is much more complex and involves correcting for unreliability in predictor scores before making the other corrections, then adjusting the resulting validity coefficient to re-introduce measurement error in the predictor. Given the limited evidence for indirect range restriction in most cases,[55] this procedure is not recommended in situations that require a great deal of assumptions to be made in calculations.

Generalizability

A single validation study is usually not sufficient for understanding the degree to which the reported validity will generalize to a new situation. The results of one study should be corroborated in at least one other validation study using similar measures with large samples. It is possible to compute an approximate confidence interval around validity estimates if it is not reported. The smaller the confidence interval, the more likely the validity coefficient will generalize to other similar settings.

The report of a meta-analysis should include the 80% credibility interval around the validity estimate. If this interval is large, it probably indicates the presence of moderator variables that influence the level of validity. If moderator analyses were not performed, ask about the potential moderators and ask to see the studies that are most similar to your situation. Highly variable results across studies make it less likely that the mean validity coefficient will generalize to a new setting.

Sometimes scores from several different assessments are weighted and combined to create a single score that is used for selection, or selected scales, from one assessment are weighted and combined in the same way. If the weights are based on data from a single study (empirical or statistical weights), they should be applied, and the validity coefficient computed, in an independent sample to determine the extent to which those weights generalize to another situation. This process is called ***cross-validation***.

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

[55] Sackett et al. (2022)

Practical Recommendations for Correcting Validity Coefficients

This report has reviewed recent research on corrections for range restriction and criterion unreliability in primary studies and meta-analyses and provided guidelines for evaluating reports supporting the validity of assessments for selection. In this section, we summarize this content in a concise set of statements on statistical corrections to validity coefficients.

1. Statistical corrections are complex but generally worthwhile.

Research has consistently shown that corrections for range restriction and criterion unreliability yield less biased estimates of population validity coefficients.[56] The key is to make sure that all steps are completed properly and any calculated or estimated parameters for the formulas are reasonable and appropriate. This is essential for evaluating competing claims of validity.

2. Correct for criterion unreliability.

Correcting for criterion unreliability has long been considered a best practice when estimating validity coefficients. Use the appropriate type of reliability coefficient for the data available and have a sound basis for the value used when estimating reliability based on other data.

3. Correcting for range restriction should be done rarely.

Other than straightforward predictive validity studies in which accurate values for direct range restriction corrections are available, meaningful corrections for range restriction are usually very difficult to make in practice. Sackett et al. (2022) demonstrated that indirect range restriction corrections often overestimate validity and make very little difference when realistic parameter estimates are used.

4. Always report both corrected and uncorrected values.

If any corrections are done, the uncorrected value must also be reported, along with all necessary information to understand how the corrections were applied.

5. Err on the side of being conservative.

If values for correction formulas cannot be computed directly from the data and estimates must be made, be conservative in the estimates. Mean values obtained from other studies will not necessarily be representative of your situation. It is better to underestimate the population validity coefficient than to present a value that cannot be supported.

[56] Hunter et al. (2006)

shl.com