Technical Affairs

By Mike Aamodt, Associate Editor

Beauty May Be in the Eye of the Beholder, But Is the Same True of a Validity Coefficient?

It was one of those strange days when a series of events gets you thinking. I was reading a summary of a legal case in which a judge indicated that validity coefficients need to be at least .30 to be of value. A few hours later, I was talking to a labor attorney who said the same thing. While I was trying to determine where this .30 figure came from, I spoke with a client who said that the OFCCP didn't think the validity of their tests was "high enough." Like an Abba song that sticks in your mind, I kept thinking about the notion of what is an acceptable validity coefficient. Is it one that is statistically significant? Is there a value that all testing experts agree is minimally acceptable? Does an acceptable validity coefficient depend on the degree of adverse impact? Does it depend on the degree of utility, and if so, how much utility is enough?

Table 1. List of Participating Assessment Experts

Testing Expert	Organization
Bryan Baldwin	Assessment Consultant, Washington Department of Personnel
Dennis Doverspike	Professor, University of Akron
Eric Dunleavy	Senior Consultant, DCI Consulting
Cassi Fields	President, Fields Consulting
Art Gutman	Professor, Florida Institute of Technology
David Hamill	Consultant, Previsor, Inc.
Seth Kaplan	Assistant Professor, George Mason University
Phil Roth	Professor, Clemson University
Lance Seberhagen	Director, Seberhagen & Associates
Trevor Self	Human Resources Officer, International Monetary Fund
Charley Sproule	Director, Sproule & Associates
Joel Wiesen	Director, Applied Personnel Research
Don Zink	Attorney, Personnel Management Decisions

So that I could get my mind away from these questions and back to Dancing Queen and Waterloo, I thought it would be useful to survey some testing experts to get their thoughts about what it takes for a validity coefficient to be acceptable. A list of the experts who responded can be found in Table 1. Not surprisingly, the consensus of the experts was that I had asked some interesting, but difficult, questions.

Art Gutman and **Eric Dunleavy** provided some insight into the source of that .30 rule of thumb. **Eric Dunleavy** thought the source might be the 1992 Psychological Bulletin article by **Jacob Cohen**. **Art Gutman** concurred and provided the following legal thoughts:

> *What sticks out to me about r = .30 is* ***U.S. v. Delaware*** *(2004). Dick Jeanneret and Harold Goldstein testified in that case. Goldstein testified that based on Cohen, correlation coefficients in a range of .10 to .20 were considered 'low to moderate' and the judge wrote the following footnote:*
>
> > *"25 At trial, Dr. Goldstein testified that this range, from .10 to .20, was the "low to moderate" range of correlations. (Tr. Vol. 5, 1413:1-24; 11414:19-1415:3.) However, as discussed infra at n. 28, the statistical text cited at trial states that correlations of .1 are described as low and correlations of .3 are described as moderate. (See Tr. Vol. 2, 402:4-11; Tr. Vol. 3, 654:9-655:4; Tr. Vol. 4, 1042:11-1043:10 (citing a standard statistical text by Dr. Cohen).) Because none of the correlations developed by Dr. Jeanneret are .3 or greater, Dr. Goldstein's testimony is unpersuasive to the extent that it characterizes the range of correlations as moderate. (See Tr. Vol. 5, 1414:19-1415:3.)"*

Question #1: Is there a minimum value for a validity coefficient that would generally be accepted by testing experts? If so, what is it?

The overwhelming response to this question was, "no." The consensus of the experts was that:

- Validity coefficients are on a continuous scale and that higher coefficients are more valuable than lower coefficients.

(continued on next page)

- A coefficient would need to be statistically significant to even be considered as potentially useful.
- Whether a validity coefficient is "good enough" would depend on the potential utility of the test. (For an explanation of utility, refer to the technical affairs column in the February, 1999, issue of the Assessment Council News http://www.ipmaac.org/acn/feb99/techaff.html).

Perhaps my questions to the experts should have addressed the issue of how much utility (e.g., percentage increase in successful employees, monetary savings) is enough to establish job relatedness. That is, would an annual savings of $1,000 per employee be enough to establish job relatedness? Would a 5% increase in successful employees be enough?

Comments by Assessment Experts

Bryan Baldwin

Conducting these types of studies is generally so challenging for organizations, particularly in the public sector, with so many decisions and assumptions that need to be made, that choosing an "acceptable" level of a validity coefficient seems silly. Assuming the study is of high quality, I would be more interested in the practical implications of the coefficient rather than the number in isolation.

Lance Seberhagen

Some people think that any validity coefficient is acceptable, even an "Irish" coefficient[1] (.07, .08, etc.), as long as it is statistically significant, regardless of whether the correlation was inflated due to capitalization on chance. This view is incorrect.

Under professional standards (APA, SIOP), the evaluation of test score interpretations is based on validity evidence from all sources, not just a single validity coefficient, and there are no absolute standards for a minimum acceptable validity coefficient. Professional judgment is needed to evaluate each situation. Within that framework, the following factors are normally considered when evaluating a criterion-related validity coefficient:

- *The uncorrected correlation should be statistically significant (two-tailed test, .05 level of statistical significance).*
- *Appropriate steps (e.g., cross-validation, Bonferroni corrections should be taken to avoid inflated correlations due to capitalization on chance).*
- *The correlation should be large enough to achieve a practical purpose.*
- *The criterion measure should provide a fair and reliable assessment of factors that are important to job performance or other legitimate business objectives (e.g., safety, customer service).*

Textbooks on testing (e.g., Cronbach, 1990, p. 167) have traditionally given .20 as a guideline for a minimum validity coefficient to have practical value. This level of validity is consistent with the following "rough guide" submitted by SIOP (aka, APA Div. 14) to the court as part of an amicus brief in ***U.S. v. Georgia Power*** *(5th Circuit, 1973): "A correlation between a single employment test and a measure of job performance of approximately .20 is often high enough to be useful, and such correlations rarely exceed .50...." (quoted in Lindemann, Grossman, & Cane, 1996).*

The federal government has taken a position on the issue in the U.S. Department of Labor's (DOL) guidebook entitled ***Testing and Assessment: An Employer's Guide to Good Practices*** *(2000 – access via http://www.onetcenter.org/dl_files/empTestAsse.pdf). DOL's guidebook (p. 30) provides these guidelines for interpreting validity coefficients:*

- *Above .35 = Very beneficial.*
- *.21 - .35 = Likely to be useful.*
- *.11 - .20 = Depends on circumstances.*
- *Below .11 = Unlikely to be useful.*

The ***Uniform Guidelines*** *(Sect. 14B-6) says that (1) there is no minimum acceptable validity coefficient for all employment situations, (2) validity should be evaluated for each case, and (3) the greater the adverse impact, the greater the validity must be to justify the test. The courts have ruled in a number of cases that the validity coefficient must be at least .30 if the test has an adverse impact – see* ***Boston Chapter, NAACP v. Beecher*** *(1st Cir, 1974),* ***Ensley Branch, NAACP v. Seibels*** *(5th Cir, 1980),* ***EEOC v. Atlas Paper*** *(6th Cir, 1989).*

Dennis Doverspike

... the real issue is utility. In some cases, even a very small validity coefficient could lead to substantial utility. Besides, the tests with the worst validities, personality inventories, are those that are most likely to be argued for by the same people arguing that there should be a minimum validity. I am assuming you want us to respond as testing experts, as scientists involved in the science of testing, and not in terms of the law.

David Hamill

This is a great question. We always talk about validity in terms of "degrees of validity" or evidence of validity, and not necessarily as "invalid" or "valid." There is also the practice to compare estimated validities from one assessment to another to make the conclusion if a test is valid or not. For instance, cognitive ability tests are more valid than unstructured interviews. This question could be addressed using the Angoff method. What is the lowest

(continued on next page)

validity coefficient that a test can have to still be considered valid? My guess is that it would fall between unstructured interviews (invalid) and a T&E rating (just barely valid). ~.11?

Seth Kaplan

I am not sure what testing experts generally think about this question. In my own opinion, asking what would be the weakest acceptable validity coefficient is not a useful question to ask. The answer depends on a host of factors such as the degree of adverse impact, the cost and utility of alternative procedures, the nature of the criterion, the personal objectives of management (e.g., increasing diversity), and so forth. There is not, or at least should not be, a "one size fits all" decision-rule. If we are selecting nuclear power plant operators versus restaurant servers, what is acceptable and useful likely differs. Simply examining statistical results is a narrow view of validity. Practical and social concerns must be considered as well. I worry that what is considered an acceptable minimum value derives primarily from court decisions. Our job is to consider all relevant factors and to make the courts aware of these.

Charley Sproule

I believe that the U.S. Department of Labor's guidance is reasonable, but feel that determining what is a minimum value for a validity coefficient is a matter of professional judgment which depends upon the type and nature of the assessment device(s), evidence from previous research for the occupation and type of assessment device(s), and the context or situation (e.g., the variability in the ability level and relevant preparation of the candidate group, hard vs. soft criterion measure, selection ratio, likely utility, etc.). A standard of meeting statistical significance standards and having practical value appears most appropriate to me. In my view, the zero-order coefficient would need to be at least statistically significant at the .05 level, and with a large sample I would want a minimum coefficient at least in the mid-teens (e.g., .15).

Generally, for assessment devices where research typically finds high average validity (structured interviews, job knowledge tests, behavioral consistency T&E's, work sample tests) I would want a minimum zero-order coefficient of .20. For assessment devices where the research literature has typically found low validity (e.g., point method ratings of training and experience) I would want a minimum zero-order coefficient of .15; however, I do not recommend such assessment methods.

Joel Wiesen

I speak for myself here, not for testing experts in general. Any non-zero validity coefficient has potential utility, depending on the selection ratio and on the variability of ability of applicants, so there is no lower, non-zero limit for a validity coefficient that is operative in all settings. Of course, the observed validity itself is influenced by the variability of job applicants' ability, so the validity of a test with a new applicant pool may differ from past observed validity. Given this, the lower limit of a validity coefficient for a selection procedure may be evaluated jointly with utility (or expected utility), practicality (e.g., testing time), and legality. Of course, when considering possible alternative selection procedures, validity should be one of the decision criteria.

Don Zink

Courts have in some instances, however, accepted correlations in the .20s as sufficient to justify the use of a particular selection procedure. Those situations are most likely to arise in selecting for positions involving safety or critical skills. For example, in ***Spurlock v. United Airlines*** *(college degree required; 10th Cir. 1972) the court wrote:*

> *"[Contrasted with jobs requiring a small amount of skill and training] when the job clearly requires a high degree of skill and the economic and human risks involved in hiring an unqualified applicant are great, the employer bears a ... lighter burden to show that his employment criteria are job-related. ... The courts, therefore, should proceed with great caution before requiring an employer to lower his pre-employment standard for such a job."*

*In another case, involving police officers, a correlation of .21 (after correction for restriction of range) was accepted as establishing predictive validity of a promotion examination. (****Pennsylvania v. O'Neill,*** *(E.D. Pa., 1979). Age limits commonly are accepted in several contexts, e.g., bus drivers. (See* ***Usery v. Tamiami Trail Tours****, 5th Cir.1976).*

The experts provided some written sources in which a testing expert or a judge (not the same thing) tried to identify the point at which a correlation coefficient becomes useful. These are shown in Table 2. When looking at these coefficients, keep in mind that some of these opinions are based on correlation coefficients in general rather than validity coefficients. Thus, the courts' reliance on Cohen's estimates are probably misguided as Cohen was referring to correlations in general rather than to validity coefficients.

What can we conclude from the experts? All other things being equal, validity coefficients above .30 are probably going to be acceptable evidence of validity whereas correlations below .11 are not. The battleground seems to be validity coefficients that fall between the range of .11 and .29. If one were to use the Department of Labor guide as well as the court's decision in Pennsylvania v. O'Neill (1979), .21 seems to be the tipping point.

Interestingly, in U.S. v. Delaware (2004), although the court referred to Cohen's .30 rule-of-thumb, and ultimately

(continued on next page)

ruled that the passing score for the test was too high, it did find the level of test validity as being acceptable. The uncorrected correlation between the test score and a composite measure of performance was .24 (.37 corrected for range restriction and criterion unreliability). The court noted, "The evidence demonstrates that the relationship between Alert scores and performance in the relevant areas of the Trooper job is relatively weak but still provides an appropriate basis for decision-making by the State. In other words, the Alert has generally low criterion validity but its predictive power is statistically significant (Note 42)."

Question #2: Is there a *minimum* value for a reliability coefficient that would generally be accepted by testing experts? If so, what is it?

The experts expressed some interesting views on this topic. Although most texts indicate that .70 is the standard for reliability, there are certainly different viewpoints. **Eric Dunleavy** mentioned that in a case with which he is familiar, an expert testified that a test must have a reliability of at least .90 to be considered reliable! Contrast this with the statements

Table 2. Summary of opinions regarding the usefulness of validity coefficients

Source	Correlation	Comment	Basis of Estimate
Cohen (1992)	.50	Large correlation	Correlation coefficients in general
Department of Labor (2000)	.36	Very beneficial	Validity coefficients
Cohen (1992)	.30	Medium correlation	Correlation coefficients in general
NAACP v. Beecher (1974) NAACP v. Seibels (1977) EEOC v. Atlas Paper (1989)	.30		Correlation /validity coefficients
Clady v. County of Los Angeles (1985)	.30	Courts generally accept correlation coefficients above +.30 as reliable (sic)	
Zamlen v. City of Cleveland (1988)	.30	Correlation coefficients of .30 or greater are considered high by industrial psychologists.	
Spurlock v. United Airlines (1972)	.20s	Court opined that employer bears a lighter burden "when the job clearly involves high degree of skill and the economic and human risks are great."	Validity coefficients
U.S. v. Delaware (2004)	.24	Court noted, "…the relationship between Alert scores and performance in the relevant areas of the Trooper job is relatively weak but still provides an appropriate basis for decision-making by the State."	Validity coefficients
Department of Labor (2000)	.21	Likely to be useful	Validity coefficients
Pennsylvania v. O'Neill (1979)	.21	Court accepted corrected coefficient of .21 as demonstrating validity	Validity coefficient corrected for restriction of range
Lindermann & Grossman (1996)	.20	High enough to be useful	Validity coefficients
Cronbach (1990)	.20	Sometimes makes an appreciable practical contribution (p. 167)	Validity coefficients
Anastasi (1988)	.20	May justify inclusion of a test in a selection program (page 169)	Validity coefficients
Cohen (1992)	.10	Small correlation	Correlation coefficients in general
Department of Labor (2000)	.10	Unlikely to be useful	Validity coefficients

(continued on next page)

of **Dennis Doverspike** and **Joel Wiesen** who essentially argued that if a test demonstrates sufficient criterion validity, its reliability is not important. Several of the experts suggested that the answer to this question may depend on the type of reliability used as well as the way in which the test score will be used.

Seth Kaplan cited the article by Lance, Butts, and Michels (2006) that discusses the origin of the .70 rule of thumb often attributed to Nunnally (1978). I had not read this article before and it is an eye opener about the truth behind four psychometric rules of thumb that we all "know to be true." I highly recommend the article.

Comments by Assessment Experts

Dennis Doverspike

No. Most tests used in the public sector are multidimensional. So reliability in this context makes no sense anyway. Many other tests, such as biodata, are designed to not be reliable. Besides, again the lowest reliabilities will probably be found for personality tests. But for many tests in the public sector, the importance of reliability is debatable. But that is a whole article in itself.

Eric Dunleavy

I don't think rules of thumb are particularly useful when it comes to a minimally acceptable reliability coefficient, mostly because some constructs are more difficult to measure than others, and because the type of reliability coefficient matters. For example, a construct that is more difficult to measure than another construct may explain variance in an outcome that other constructs do not. Most people familiar with personnel psychology are aware that reliability provides an upper bound for a correlation in the criterion-related validity case. However, it may be difficult to understand how reliability actually influences validity and selection decisions in a particular construct.

One practical way to consider whether test reliability is 'minimally acceptable' is via the standard error of measure (SEM), which captures the amount of error in test score units. For example, given reliability for a test, would you feel comfortable adjusting a cut score down by two SEMs? Would you feel comfortable using a banding technique for selection decisions where the bands were grouped by the SEM? In this context, an unreliable measure will have a large SEM, and as such, adjusting a cut score or a banding strategy via the SEM may have important consequences for how many candidates pass a test or get selected. If pass/fail or selection decisions are very different when considering and not considering the SEM, do you feel good about making any decision based on that test?

Cassi Fields

*In my most recent experience in testing litigation (**Dudley, et al. and Wood, et al. v. City of Macon, GA,** ruling: September 19, 2006), experts argued over the level of acceptable reliability (r_{xx}) of public safety tests. Two main sources were cited and debated – Schmitt's 1996 article, "Uses and Abuses of Coefficient Alpha" (Psychological Assessment, 350-353), and Gatewood and Feild's (2001) "Human Resource Selection" textbook, specifically the section called 'How High Should a Reliability Coefficient Be?' (144-146). Both discussions on reliability are highly informative and should be read by all interested in this topic.*

Schmitt (1996) said that researchers often consider .70 an acceptable reliability, but goes on to give an in-depth explanation of why this one value is shortsighted. His opinion is that a lack of understanding of reliability (particularly coefficient alpha) has led to misuse and misinterpretation of reliability and validity data. Gatewood and Feild (2001) said reliability is crucially important when precise measurement is needed (e.g., in cases of selection) and should have a minimum value of .85. Essentially, Gatewood and Feild (2001) claim that error variance may unfairly inflate or deflate test scores. Their requirement for very high reliabilities focuses on the need for "precision" in measurement. I understand their argument to mean that a measure must get very close to measuring the test taker's true score, with little to no error, or it is not good for making high-stakes decisions. In my opinion, this argument overly limits the use of predictors.

Maybe the question of what is an acceptable reliability value is a misdirected question. The better question is 'in each individual situation, did the reliability of the predictor overly restrict the potential validity of the measure?' Reliability estimates of a predictor as low as .50 - .60 might still be statistically valid, just less valid than a predictor with higher reliability. A measure with restricted validity still accounts for variance in job performance, and therefore has value. The question of how much different levels of reliability restrict validity might have a solid statistical and rational answer in a criterion-related study where there is a validity coefficient, but it cannot be answered in a content-related validation study.
This debate also ignores the question of whether the reliability estimate of each test component in a test composite must be .70, .85 or higher, or if it is acceptable for the linear composite to reach these values when the test component reliabilities are combined. If none of the test components serve as hurdles, then it seems the test composite reliability should be examined and not the individual test component reliabilities.

(continued on next page)

When the courts are presented with values and rules by testing experts, it unnecessarily handcuffs some industries. In my professional specialty, public safety testing, the public safety organizations I work with are forced to proceed with traditional testing measures and classical true score test statistics since that is what the legal system (attorneys and judges) understands. This precedent prevents public safety organizations from inventing new selection systems and measures since the court, by definition, relies on precedent. Industries that are not so controlled by employment litigation have the luxury of trying non-traditional assessments such as video situational judgment tests, on-line selection, virtual job tryouts and other unique selection programs because they are not so severely limited by traditional testing rules including one specific acceptable reliability value.

I think it is not in the best interest of our profession to set one reliability value to use as the "go by."

Seth Kaplan

Again, I am not sure what testing experts consider acceptable. I would imagine somewhere in the neighborhood of .70 to .80 – although these rules of thumb are not founded in the literature (Lance, Butts, & Michels, 2006). Here too, however, various factors such as the variability in the sample, the nature of the construct, number of items, etc. are relevant. I also would mention that the impact of reliability on a validity coefficient is not the same as the impact on individual decisions. One could have a reasonable validity coefficient and a lower reliability estimate (e.g., .60). The measure may be predictive in the aggregate, but using it to make decisions about individual respondents' that affect their life outcomes would be inappropriate.

Lance Seberhagen

Reliability is the degree to which a test is free from random error. A reliability coefficient is an estimate of the percentage of systematic (nonrandom) variation in test scores. Reliability is a necessary but not sufficient condition for validity. The square root of the reliability coefficient provides the upper limit of the validity coefficient.

Professional standards (APA, SIOP) do not provide a minimum acceptable reliability coefficient. Instead, professional standards say that professional judgment should be used to evaluate each situation. There are various methods (or models) for estimating reliability (e.g., test retest, internal consistency). The method used to estimate reliability should be relevant to the way in which the test scores are used.

*Textbooks on testing (e.g., Nunnally, J.C., **Psychometric Theory**, 2nd ed., 1978) generally say that a reliability of .70 is sufficient for most uses, but reliability coefficients as low as .50 may be acceptable (Nunnally, J.C., **Psychometric Theory**, 1967; Thompson, B., **Score Reliability**, 2003).*

*The U.S. Department of Labor has published a guidebook on **Testing and Assessment: An Employer's Guide to Good Practices**, 2000. The guidebook (p. 23) provides these guidelines for interpreting reliability coefficients:*

- *.90 – up = Excellent.*
- *.80 - .89 = Good.*
- *.70 - .79 = Adequate.*
- *Below .70 = May have limited applicability.*

Charley Sproule

I believe that .70 is the minimum desired reliability (i.e., internal consistency or test-retest reliability) testing experts would accept for paper-and-pencil tests; and .60 would be the minimum desired reliability for rating-based assessments (e.g., structured interviews) since .60 is the average reliability of performance evaluations, which are often used as criterion measures.

Joel Wiesen

I speak for myself here, not for testing experts in general. There are various measures of reliability (e.g., internal consistency and test-retest), and they measure somewhat different things. Here I speak of test-retest reliability. Reliability may be evaluated in at least two ways: (a) how it affects the psychometric characteristics of the selection process, and (b) how it may affect individual applicants. From a psychometric viewpoint, test reliability contributes to test validity, since an unreliable test can have no validity. So, in this respect, if we know the validity we need not consider the reliability. From the viewpoint of individual applicants, the lower the test reliability the more the scores may be due to non-merit factors (chance or various sources of systematic error). In my expert witness work, I have seen changes of 20+ points (on a 100-point scale) on rescoring an assessment center after an appeal. Large changes on regrading give rise to applicant perception that the assessment process is capricious. We should strive for high reliability, and should try to measure reliability, when feasible and practical. Of course, when considering possible alternative selection procedures, reliability may be one of the decision criteria.

(continued on next page)

Question #3: In your opinion, what is the lowest uncorrected validity coefficient that you believe would indicate that an inference from a test has acceptable criterion validity? That is, if a test had adverse impact, what is the magnitude of the validity coefficient that I would need for you to acknowledge that the test is job-related?

Expert opinion was divided on this question. Some experts argued that the presence of adverse impact is irrelevant because a test is either job-related or it is not. Others argued that the presence of adverse impact (or any social consequence) would necessitate a higher validity coefficient than in a situation without social consequences.

Comments by Assessment Experts

Dennis Doverspike

Why does it matter if the test has adverse impact? If you are talking from a scientific or expert perspective, why does the adverse impact of the test matter? Either we are talking expertise and science here or practical legal issues. But those are different questions. So, from a scientific or expert point of view, adverse impact is not relevant to the discussion. Besides, again, those tests with the lowest validities are likely to be personality tests, which are the tests that plaintiff's experts are likely to argue for. Again, from a scientific standpoint, you would want to argue the real issue is utility. Now I would certainly argue that adverse impact does enter into the utility perspective. But the costs of hiring a bad janitor, a bad police officer, or a bad fighter pilot, are not equal.

Lance Seberhagen

It depends on the situation and the alternatives available. Given caveats in my answer for validity coefficients above, an uncorrected validity coefficient of .30 would be generally acceptable, .20-.29 would be borderline, and less than .20 would be suspect.

Trevor Self

The issue here is not with the job-relation between the test and some measure of performance, which is indicated by the validity coefficient. Rather, the issue is with the construct- and content-related evidence for validity in the measure of the performance criterion. The underlying argument of the adverse impact case is that the performance measure is either an inaccurate measure of the performance construct or that the performance construct is contaminated. The adverse impact case becomes (or should become) moot (barring pre-text) if the criterion-related validity is significant and the criterion appears construct- and content-valid (with apologies to Landy, Messick, and others for reverting to the tripartite terminology). In that case, the test seems to meet the criteria for a BFOQ exemption. Without conflating the issue with adverse impact concerns, the value of the test is what it adds above baseline prediction. If your prediction is no better than random selection, then any real (i.e., significant and reliable) relation between a predictor and the criterion is worthwhile (forget, for a moment, consideration of costs of implementation and so forth). Conversely, if the job requirements are low (such that baseline success is high), a much higher criterion-related validity coefficient would be needed to substantiate the use of the instrument. It's about meaningful improvement in the prediction of performance. Sometimes even small improvements are meaningful. Of course, there are other issues, such as the number of predictors and assessment of the squared semi-partial correlation coefficient versus the bivariate correlation between the predictor and criterion. Likewise, the issue is complicated with consideration of dynamic criteria, such as prediction of both readiness for an immediate position and potential for later positions.

Rote rules and cut-offs too frequently erode value by supplanting appropriateness with convenience.

Joel Wiesen

As explained in my answer to Question #1, any non-zero validity coefficient may have a useful level of utility, depending on the situation. The question is whether there is a reasonable alternative selection procedure with less adverse impact and equal (or, perhaps, similar) validity. Random selection might be considered an alternative selection procedure, and it would be expected, on average, to have no adverse impact. We may ask, is the increase in validity (and utility) enough to justify the observed level of adverse impact? This question may be approached from practical, legal, and societal perspectives.

Question #4: What is the lowest corrected validity coefficient that you believe would indicate that an inference from a test has acceptable criterion validity?

Most of the experts chose not to answer this question in detail. My take on the experts' responses is that this is uncharted territory and the answer would depend on the magnitude of the uncorrected coefficient as well as what the coefficient was being corrected for.

(continued on next page)

Comments by Assessment Experts

Lance Seberhagen

Again, it depends on the situation and the alternatives available. I would generally want to see a corrected validity coefficient of at least .30.

Trevor Self

I wouldn't use a corrected validity coefficient, especially if courts establish different cut-offs for corrected and uncorrected coefficients. Corrected estimates are inflated due to an assumption that unreliability detracts wholly from the true relation between the predictor and criterion. This is unlikely, as is partially evident from corrected validity coefficients that exceed 1. (In some cases, MTMM studies have likewise suggested the inflation of corrections.) Not only do such inflated estimates skew diagnostics and subsequent perceptions within an organization, but plaintiffs can (rightly) also use this inflation to their advantage by claiming that the true relations are not as large as they appear.

Question #5: If a validity coefficient is statistically significant, is that enough to imply job relatedness? If not, what other factors would you consider in determining if the test is job-related and has practical significance?

I think a fair summary of the responses to this question would be:

- The validity coefficient must be statistically significant.
- Any statistically significant coefficient has the potential to be useful.
- Statistical significance by itself is not enough to imply job relatedness.

Comments by Assessment Experts

Dennis Doverspike

No. See Binning and Barrett or the APA Guidelines. It is an overall decision about the content, construct, and criterion-related validity of the test.

Trevor Self

Yes, but only to the extent that the criterion measure is valid (i.e., a relevant & complete representation of the job performance construct). Job-relatedness isn't dichotomous; there is a degree of relation to the job. If the measure of that relation is significant, there is a relation. The effect size indicates how much of a relation there is. Some other organizational measures of effectiveness (e.g., monetary value of the standard deviation of performance, ProMes, etc.) then must be used to translate that relation to estimated value for the organization.

An adequate answer to the second question would be horrendously long. There are many (often contextually dictated) considerations that affect the practical significance of test usage (e.g., the face validity of the test; the ability to integrate the test with other instruments, practices, formats, delivery systems, etc.; the degree to which the instrument aligns with the organizational culture; the degree to which the instrument will be useful long-term, in light of organizational strategy; the source of the instrument; etc.).

Charley Sproule

If a validity coefficient is statistically significant, that does not imply job relatedness. An assessment device could have criterion-related validity without being job-related. I believe that we always need to have evidence of content validity for assessment devices used in the public sector, since local criterion-related studies are usually not feasible.

I believe that job relatedness and validity are not synonymous. For example, number of years of education and certain biodata items (e.g., educational level of parents) are sometimes found to have a low positive correlation with criterion measures, but these may not be job-related measures. Job relatedness differs from content validity in my view. The following definition of "job relatedness" is from an IPMAAC Personnel Assessment Seminar:

> *JOB RELATEDNESS. A showing of a relationship between the requirements of a selection procedure and the requirements of a job such that the relation is readily apparent (e.g., a job requires reading so a reading test is given; a job requires typing so a typing test is given). This is related to content validation but lacks the complete documentation of content validation.*

Joel Wiesen

I see job related as synonymous with valid, so see my answers to the questions above. Additionally, we may want to consider how much of the job the test and the criterion cover; however this may go beyond your question.

Concluding Thoughts

So, where are we after all this expert opinion? Perhaps the same place we were before. The majority opinion seemed to be that rules-of-thumb for validity and reliability coefficients are not a good idea because there are so many factors that affect whether a test is job-related or is "reliable enough" to

(continued on next page)

produce stable results. The consensus of opinion was that validity and reliability are continuous, rather than dichotomous. Thus, there is not a place at which a test becomes "valid" or "reliable."

Unfortunately, when an organization is sued, the courts seem to treat validity as a dichotomous variable and rule whether a given validity coefficient is "high enough" to be valid. Our experts believe the courts should consider the potential utility of the test scores, but again, we have no acceptable cutoff regarding how much utility a test must have to be considered job-related.

Because of the great response to this topic, I am going to try to organize a panel discussion either at IPMAAC for 2008 or SIOP for 2009. Several of the experts in this column have agreed to participate. If any of the ACN readers have thoughts on the topics covered in this column, please email them to me at maamodt@radford.edu.

Notes

[1] Lance credits Mary Tenopyr for the term "Irish" validity coefficient. The term may date back to the 1930s or 1940s, but Lance first heard it when Mary used it at a Div. 14 (SIOP) session at an APA convention in the 1970s, probably about the time of the US v. *Georgia Power* case. Lance said that perhaps she got the term from someone else (Al Maslow, Dick Barrett, Don Grant, or Bob Guion are likely suspects), but it sounds like her sense of humor, so Lance gives her credit for it.

References

Anastasi, A. (1988). *Psychological testing* (6th edition). NY: Macmillan.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: a conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74, 478-494.

Boston Chapter, NAACP v. Beecher, No. 74-1067 (U.S. Ct. App. First Cir. 1974).

Clady v. County of Los Angeles, 770 F.2d 1421 (U.S. Ct. App. Ninth Cir. 1985).

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

Cronbach, L. J. (1990). *Essentials of psychological testing*. NY: HarperCollins.

Ensley Branch, NAACP v. Seibels, 14 FEP Cases 670 (N.D. Ala 1977).

Gatewood, R., & Feild, H. (2001). *Human resource selection* (5th edition). Orlando, FL: Harcourt Publishers.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9(2), 202-220.

Lindemann, B., Grossman, P., & Cane, P. W. (1996). *Employment discrimination law* (3rd edition). NY: BNA Books.

Nunnally, J. C. (1967) *Psychometric theory*. NY: McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory, 2nd ed.* NY: McGraw-Hill.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353.

Thompson, B. (2003). *Score reliability*. Thousand Oaks, CA: Sage Publications.

Zamlen v. City of Cleveland, 686 F. Supp. 631 (U.S. Dist. Ct. No. Dist. of Ohio, Eastern Div. 1988).

HR Humor

So you want a day off!

So you want a day off? Let's take a look at what you are asking for:

- There are 365 days this year.
- There are 52 weeks per year in which you already have 2 days off per week, leaving 261 days available for work.
- Since you spend 16 hours each day away from work, you have used up 170 days, leaving only 91 days available.
- You spend 30 minutes each day on coffee break. That accounts for 23 days each year, leaving only 68 days available.
- With a one-hour lunch period each day, you have used up another 46 days, leaving only 22 days available for work.
- You normally spend 2 days per year on sick leave. This leaves you only 20 days available for work.
- We are off for 5 holidays per year, so your available working time is down to 15 days.
- We generously give you 14 days vacation per year which leaves only one day available for work and I'll be damned if you're going to take that day off!