TEST DEVELOPMENT AND USE: NEW TWISTS ON OLD QUESTIONS

Wayne F. Cascio and Herman Aguinis

Over the past several decades, there have been some significant advances in psychological science, specifically in our knowledge about important questions to address with respect to the development and use of assessment tools. This article focuses on developments in research and guidelines for practice in five selected areas that, if applied, will lead to more informed use of assessment tools. The five areas that we discuss are validity generalization, statistical significance testing, criterion measures, cutoff scores, and cross-validation. © *2005 Wiley Periodicals, Inc.*

Consider the following scenario. You have just assumed a new position as Chief Human Resources Officer (CHRO) for a U.S.-based start-up firm that provides sales and service of sophisticated computer-based simulation software. There is growing demand in the marketplace for the products your firm sells and services. As a result, you expect to hire many new associates for a variety of sales and service jobs over the next several years.

In considering alternative selection systems for these jobs, a number of issues immediately present themselves, and you jot down some of the most prominent ones:

1. With respect to the validity of any tool or instrument that we use as a basis for choosing among candidates, can we rely on evidence developed in similar situations elsewhere, or do we have to develop our own "in-house" data as well?
2. What types of measures (i.e., criterion measures[1]) should we use to assess the performance of associates on the jobs in question?
3. If we develop "in-house" data on predictors (i.e., assessment tools) and measures of job performance on one sample of individuals, do we need to use a different sample in order to check (that is, cross-validate) the results obtained with the first sample?
4. Is evidence of statistical significance sufficient to draw conclusions about

Correspondence to: Wayne F. Cascio, The Business School, University of Colorado at Denver and Health Sciences Center, Campus Box 165, P.O. Box 173364, Denver, CO 80217-3364, (303) 556-5830, Wayne.Cascio@cudenver.edu

Human Resource Management, Fall 2005, Vol. 44, No. 3, Pp. 219–235

DOI: 10.1002/hrm.20068

the meaningfulness of "in-house" data that show relationships between scores on assessment tools and measures of performance?

5. Finally, if we set minimum cutoff scores for performance on selection procedures, what key legal and psychometric issues do we need to know about?

…over the past several decades, there have been some significant advances in psychological science, specifically in our knowledge about important questions to address with respect to the development and use of assessment tools.

To answer these questions, you turn to the literature on psychological assessment. From your reading, you learn that over the past several decades, there have been some significant advances in psychological science, specifically in our knowledge about important questions to address with respect to the development and use of assessment tools. This article is not an exhaustive review of that body of research. Nor is it a broad literature review that focuses on summarizing current knowledge and identifying fruitful areas for future research. Rather, its purpose is to inform HR practitioners of current research findings in five key areas relevant to assessment practices that they commonly confront, and to distill guidelines for practice from that body of knowledge. These areas are controversial, in that even experts disagree about appropriate methods for dealing with them, and HR practitioners may therefore differ in their approaches to them. At the same time, there is general agreement among researchers about the guidelines that we present.

At a broader level, a recent study found substantial differences between established research findings in the domain of staffing and practitioner beliefs (Rynes, Colbert, & Brown, 2002). As the authors emphasized, "lack of awareness of broad selection principles can be very costly to organizations" (p. 165). The importance of the five areas we have selected is highlighted by the inclusion of three of them (statistical significance testing, criterion measures, and cutoff scores) in a recent scientific update of the federal Uniform Guidelines on Employee Selection Procedures (Cascio & Aguinis, 2001). Those authors examined systematically each section of the 1978 Uniform Guidelines and identified several areas that require revision and update in light of subsequent scientific developments. We address two additional areas in this article: validity generalization and cross-validation. These are the same five areas that the CHRO's questions address, as cited at the beginning of this article.

In the scenario described earlier, one question that the CHRO of the start-up firm raises is whether evidence developed elsewhere can be used to support the validity of the same instrument or procedure in a different, but similar, situation. Suppose the CHRO has read meta-analytic results of validity studies and wonders if such results might apply to the situation she faces. She attended a professional conference recently, where she learned that meta-analysis, which is a quantitative summary of results from different studies on the same topic, is used for two purposes. The first is to draw more general scientific conclusions and the second is to use the results of validity evidence obtained from prior studies to support the use of a test in a new situation (Cascio & Aguinis, 2005). The latter use is termed *validity generalization* (VG). What else should the CHRO know about VG? To answer that question, let us consider some recent developments in this area.

Validity Generalization

Schmidt and Hunter (1977) hypothesized that the variability across studies in validity coefficients, even when jobs and tests appear to be similar or essentially identical, might not represent genuine differences. In developing a model to test this hypothesis, they identified seven potential reasons that might explain the variability in observed validity coefficients. The most important of these reasons is sampling error. The many VG studies and technical refinements of VG procedures that now exist in the literature in applied psychology (e.g., Aguinis, 2001; Aguinis & Whitehead, 1997; Hunter & Schmidt, 2004; Raju, Anselmi, Goodman, & Thomas, 1998; Schmidt & Hunter, 1998) suggest that VG is a robust phenomenon. Perhaps the most important implication of this work is that it has called attention to the fact that the mean of several validity coefficients may be a better

basis for inferring a valid relationship between predictor and criterion than any one coefficient (Society for Industrial and Organizational Psychology, 2003).

Legal Status of VG

A thorough search of the LexisNexis database revealed two things. One, only three cases that relied on VG have reached the appeals-court level (*Aguilera* v. *Cook County Police and Corrections Merit Board* [1985]; *Bernard* v. *Gulf Oil Corp.* [1989]; *EEOC* v. *Atlas Paper Box Co.* [1989]). Two, courts do not always accept VG evidence. In *Bernard*, for example, the court refused VG evidence by disallowing the argument that validity coefficients from two positions within the same organization indicate that the same selection battery would apply to other jobs within the company without further analysis of the other jobs. Likewise, in *Atlas Paper Box Co.* (1989), the Sixth Circuit Court of Appeals refused to accept the validity of a measure of general intelligence (the Wonderlic test) that relied on VG evidence. Atlas did no job analyses to establish the appropriateness of VG for the jobs in question. Atlas's expert never visited the company or even read the studies that formed the basis for the company's VG argument. The expert witness simply contended that the use of a measure of general intelligence is always a valid predictor. At the trial-court level, the expert was unable to support this premise when confronted with a hypothetical applicant for a firefighter position who is confined to a wheelchair and who earns the highest score on a paper and pencil test. In a concurring but separate opinion, one of the Sixth Circuit judges wrote, "As a matter of law, Hunter's validity generalization theory is totally unacceptable under relevant case law and professional standards" (p. 1501).

In reviewing the implications of this case and other VG cases, Landy (2003) noted: "When a Circuit Court of Appeals concludes *as a matter of law* that a practice is unacceptable, lawyers tend to listen . . . Further, it appears that anyone considering the possibility of invoking VG as the sole defense for a test or test type might want to seriously consider including additional defenses (e.g., transportability analyses) and would be well advised to know the essential duties of the job in question, and in its local manifestation, well." (pp. 188, 189).

…courts do not always accept VG evidence.

Guidelines for Practice

Like any other method, VG should not be used indiscriminately. In order to evaluate critically the VG evidence she reads about, the CHRO referred to in the opening vignette of this article should consider the following guidelines (presented in rough order of importance):

- In the situation the CHRO faces, provide evidence that the jobs and contexts are similar to those described in the VG study reviewed.
- Do not rely on VG as the sole basis for defending a test, if challenged. Be able to demonstrate that jobs for which a test was used are similar to the jobs included in the VG study.
- Ensure that the VG study describes clearly the aspects of behavior (in predictors and criteria) it purports to assess, along with the specific measures used to assess the strength of the relationship.
- The VG study should state that it includes all publicly available studies in the content domain of interest, not just published studies, or those that are easily available.
- The variables characterizing the studies should be selected or coded based on a priori theoretical grounds, and not just because they were available in the studies reviewed.
- Multiple raters should apply the coding scheme; measures of interrater reliability should be reported.
- VG studies should include all variables that were analyzed, including analyses of potential moderator variables, so that the CHRO can assess the extent to which chance variations in the relationships in a subset of studies might account for the results obtained.

- The characteristics of the studies included should be reported in as much detail as possible so that readers can assess the nature of the generalizations that are appropriate.

…in addition to significance testing of the correlation value or other measure of association, one also should compute a confidence interval around that single value.

Statistical Significance Testing

The CHRO described in the opening vignette questions whether tests of statistical significance can be used as a basis for conclusions about the meaningfulness of results obtained from the collection of "in-house" data. Assuming a correlation coefficient is computed to assess the relationship between scores on the new test and scores on a relevant criterion measure, the CHRO needs to interpret the importance of the resulting effect. Statistical software output indicates that the correlation coefficient is statistically significant. Is this sufficient evidence to reject the null hypothesis of no relationship between the test and the criterion and to conclude that the test is an accurate predictor of the criterion of interest? In other words, is this enough evidence to conclude that the new test is useful and, therefore, could be implemented immediately?

Null hypothesis significance testing has been, and still is, a topic of heated debate in the scientific community (e.g., Markus, 2001; Nickerson, 2000; Task Force on Statistical Inference, 2000; Tryon, 2001). Researchers have written extensively on the purpose, meaning, and use of significance testing. Some argue that significance testing is useful (e.g., Wainer, 1999), whereas others believe that it is misleading and should be discontinued (e.g., Schmidt, 1996). From the perspective of a CHRO, however, two important issues are how significance tests should be used and whether reporting significance levels is sufficient information to make a decision about using the test or whether additional information is needed.

Use of Significance Testing

Many researchers have noted that significance testing is abused and misused (e.g., Krantz, 1999; Tryon, 2001). Significance testing allows us to infer whether the null hypothesis that there is no systematic relationship between test scores and criterion scores is likely to be false. On the other hand, significance testing is used *incorrectly* when (a) conclusions are made regarding the magnitude of the relationship in the sample (e.g., a statistically significant result at the .01 level is interpreted as a larger difference than a result at the .05 level), and (b) failure to reject the null hypothesis is interpreted as evidence of a lack of relationship in the population. (In fact, a failure to detect differences in the sample may be due to insufficient statistical power.)

Guidelines for Practice

The above discussion of recent research on statistical significance testing leads to the following recommendations to address the specific questions the HR practitioner is facing:

- Reporting significance levels is important and usually welcome by the courts. However, in addition to significance testing of the correlation value or other measure of association, one also should compute a confidence interval around that single value. The confidence interval indicates the likelihood (e.g., .95, .99) that the estimated population correlation falls within the computed interval (with repeated sampling under the same conditions). The confidence interval provides a range of possible values that suggests the degree of practical significance of the effect.
- The CHRO has the choice to report the single value (and corresponding confidence interval) based on the observed validity coefficient, or a "corrected" validity coefficient (i.e., based on corrections for measurement error in the criterion and range restriction on the predictor; Hunter & Schmidt, 2004). Use the observed correlation if the objective is to understand predictive validity evidence for a specific predictor-criterion relationship (after all, the same

fallible test would be used for prediction purposes). Alternatively, use the corrected correlation if the goal to understand the more general relationship between the constructs underlying the test and the criterion (i.e., relationships between constructs are best understood when measurement error, range restriction, and other methodological and statistical artifacts are corrected for).

- Be very clear about what information the statistical significance test provides and what it does not provide (e.g., strength of the effect). Krantz (1999) and Nickerson (2000) provide more detailed information on this issue.
- Are there situations where one would consider a correlation coefficient to be too small even if it is statistically significant and its confidence interval excludes zero? To answer this question, we need to consider the fact that the validity coefficient is only one of several variables that determine the utility of the test. Other variables include the selection ratio and the cost of testing (Cascio, 2000). Thus, the size of the validity coefficient is just one of several variables to consider before deciding whether a test should be used. Considered in isolation, however, experts have argued that a validity coefficient of .30 or larger should be taken seriously (Hartigan & Wigdor, 1989). The value of .30 corresponds to what Cohen (1988) has defined as a "medium effect size" for the correlation coefficient in social science research in general. However, a correlation coefficient smaller than .30 can yield considerable utility under the right conditions (e.g., low testing cost, large number of applicants). In short, the size of the validity coefficient is just one of several factors that need to be considered in deciding whether a test should be implemented.
- Compute statistical power to rule out the possibility that the sample size was too small to detect an effect (a deviation from the population value) that actually was present.

Measures of Performance (Criteria)

The development of criteria that are adequate and appropriate is an ongoing challenge, for we know that criteria are dynamic, multidimensional, and appropriate for different purposes.

We noted earlier that a key question facing the new CHRO is to identify the types of measures (criteria) that should be used to assess the performance of associates on the jobs in question. The development of criteria that are adequate and appropriate is an ongoing challenge, for we know that criteria are dynamic, multidimensional, and appropriate for different purposes.

Criteria are critically important considerations in evaluating the usefulness of assessment tools. With respect to construct validation, efforts to validate measures of sales helpfulness or sales aptitude actually revolve around two issues: (1) *what* a test or other procedure measures (that is, the hypothesized underlying trait or construct) and (2) *how well* it measures (that is, the relationship between scores from the procedure and some external criterion measure) (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

The primary standard for choosing a criterion is relevance. In other words, if the criterion measures used are deficient (i.e., important behaviors and outcomes are not included in the measure) or contaminated (i.e., irrelevant behaviors and outcomes are included in the measure), the results of any study that uses these measures will not provide useful information. Researchers have investigated the following phenomena that have important implications for the selection of criterion measures and the conduct of studies that attempt to relate scores on assessment tools to them: (a) dynamism of criteria, (b) distinction between typical and maximum performance, and (c) multidimensionality of criteria. We discuss each of these issues next.

Dynamism of Criteria

If the rank ordering of individuals on a criterion changes over time, future performance

becomes a moving target. Under those circumstances, it becomes progressively more difficult to predict performance accurately the farther out in time from the original assessment (Keil & Cortina, 2001). Are criteria really dynamic? In other words, do performance levels show systematic fluctuations across individuals over time? The answer seems to be in the affirmative (Deadrick & Madigan, 1990; Hofmann, Jacobs, & Baratta, 1993; Hulin, Henry, & Noon, 1990). In fact, Keil and Cortina (2001) concluded that the deterioration of validity over time is a ubiquitous phenomenon.

Evidence indicates that measures of maximum performance (i.e., what employees* can *do) correlate only slightly with measures of typical performance (i.e., what employees* will *do).

Distinction Between Typical and Maximum Performance

A second issue regarding criterion measures is typical versus maximum performance. Evidence indicates that measures of maximum performance (i.e., what employees *can* do) correlate only slightly with measures of typical performance (i.e., what employees *will* do) (DuBois, Sackett, Zedeck, & Fogli, 1993; Sackett, Zedeck, & Fogli, 1988). This consideration is critical to the design of validation research. HR practitioners must determine whether the objectives of a validation study dictate a focus on typical or maximum performance.

Multidimensionality of Criteria

A final consideration regarding criterion measures is the multidimensionality of criteria. Researchers have long recognized that job performance is a multidimensional construct (e.g., Schmidt & Kaplan, 1971). Consequently, criterion measures ought also to be multidimensional (Campbell, McCloy, Oppler, & Sager, 1993).

Borman and Motowidlo (1993, 1997) and Coleman and Borman (2000) proposed a two-dimensional taxonomy: task performance and contextual performance. Task performance includes activities that are directly related to the job, including the transformation of materials into the products and services rendered by the organization, the distribution of the product, coordination and supervision of activities, and so forth (Johnson, 2001). Contextual performance (also referred to as organizational citizenship performance; Organ, 1997) is defined as those behaviors that contribute to the organization's effectiveness by providing a good environment in which task performance can occur (e.g., exerting extra effort as necessary, volunteering, cooperating). Contextual performance often is part of measures of interpersonal skills such as teamwork, cooperation, and collaboration. To the extent that such measures of interpersonal skills are part of an organization's performance management system, they are likely to be used as criteria in validation research.

Guidelines for Practice

The dynamism of criteria has the following implications for practice:

- In general, HR practitioners should attempt to identify and understand the variables that cause performance to change over time. For instance, some individuals may learn faster than others, and individuals may differ in self-efficacy, need for achievement, or self-esteem. In addition, changes in the nature of the job (e.g., from an emphasis on the product to an emphasis on customer service) may interact with individual characteristics and also are likely to affect the long-term validity of predictors. A careful analysis of the impact of each of these individual-differences variables in specific contexts will allow for the development of better criterion measures.
- Some types of predictors are more likely to predict criteria for longer periods of time than are others. Specifically, general cognitive ability tests maintain their predictive power longer than more narrowly defined ability measures (Steele-Johnson, Osburn, & Pieper, 2000). So, if long-term prediction is the objective, the CHRO in the opening vignette should use a test of general as opposed to narrow ability.

- Even though tests of general abilities are more likely to maintain their long-term usefulness, recognize that changes in the consistency and complexity of tasks are likely to diminish their validity over time (Steele-Johnson, Beauregard, Hoover, & Schmidt, 2000).
- Be aware that in some organizational contexts (e.g., service industry, fast-paced work environments), task performance is more likely to change over time (e.g., the introduction of new products and services requiring the implementation of new processes), whereas contextual performance is more likely to remain stable over time (i.e., in spite of the introduction of new products and services, it is still important to continue to provide an environment in which task performance can occur). Consequently, predictors of contextual performance, more so than those of task performance, are likely to be valid for longer periods of time.

The distinction between typical and maximum performance has the following implications for the use of criterion measures in general:

- The choice of a specific criterion measure in a validation study needs to consider whether scores are likely to be predicted by a selection procedure targeting typical or maximum performance. Selection procedures are commonly administered in environments conducive to maximum performance (i.e., applicants are aware their performance is being monitored and the assessment of performance takes place over a short period of time). On the other hand, criterion measures are commonly administered in environments conducive to typical performance (e.g., employees are not always aware that their performance is being observed, and supervisors observe job-related behaviors over a long period of time). Thus, a lack of congruence between the performance construct assessed by selection procedures (i.e., maximum performance) and the performance construct assessed by criterion measures (i.e., typical performance) may prevent the development of tests showing high predictive validity.
- The focus of a validation study will be determined in part by whether it includes a measure of typical or maximum performance as a criterion. If a measure of typical performance is included, the focus of the validation study is whether a new test can predict what employees will do. On the other hand, if a measure of maximum performance is included, the focus of the validation study is whether a new test can predict what employees can do.

Be aware that in some organizational contexts task performance is more likely to change over time whereas contextual performance is more likely to remain stable over time.

Finally, developments in the area of criterion multidimensionality lead to the following guidelines:

- When relevant to the job in question, criteria in conducting a validation study should include both task-specific and non-task-specific dimensions. In fact, in today's education and work environments where technology requires constant learning of new tools in a cooperative context and changes in organizational structure require the ability to work in teams, it could be argued that non-task-specific performance may be at least as important as task-related performance.

The two remaining issues that we shall discuss—cutoff scores and cross-validation—are both relevant to the vignette at the beginning of this article. That vignette describes the task facing the CHRO who must identify minimum levels of success on the measures used. Cutoff scores are key considerations in that effort. We address that issue next.

Cutoff Scores

In some instances, cutoff scores need not be set … in others, rules established by various governing bodies might require that a cutoff score be established.

In some instances, cutoff scores need not be set, as when rank-order (top-down) selection is used. In others, rules established by various governing bodies (e.g., at the state or local levels) might require that a cutoff score be established in order to determine which examinees will be licensed, credentialed, promoted, or graduated (Plake & Hambleton, 2001).

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) address issues of cutoff scores in several sections. For example, "Cut scores may be established to select a specified number of examinees (e.g., to fill existing vacancies), in which case little further documentation may be needed concerning the specific question of how the cut scores are established, though attention should be paid to the legal requirements that may apply" (Standard 4.19). Standard 4.21 notes: "Cut scores are sometimes based on judgments about the adequacy of item or test performances . . . or performance levels (e.g., the level that would characterize a borderline examinee). The procedures used to elicit such judgments should result in reasonable, defensible standards that accurately reflect the judges' values and intentions."

Setting Minimum Standards

When items are presented in a multiple-choice format, it is common to follow the Angoff (1971) procedure to set minimum standards. In this approach, expert judges rate each item in terms of the probability that a barely or minimally competent person would answer the item correctly. The probabilities (or proportions) are then averaged for each item across judges to yield item cutoff scores, and item cutoff scores are summed to yield a test cutoff score. The method is easy to administer, it is as reliable as other judgmental methods for setting cutoff scores, and it has intuitive appeal because expert judges (rather than a consultant) use their knowledge and experience to help determine minimum performance standards (Cascio, Alexander, & Barrett, 1988).

When items are presented in a constructed-response format, however, they permit examinees to produce a response in their own words. Such open-ended questions may be oral responses, written essays, or observation of performance by scorers (Plake & Hambleton, 2001). One approach for setting cutoff scores in such situations is termed the analytical judgment method (AJM). Panelists are given a set of examinee work samples for each of the questions comprising the assessment. Each panelist first rates each examinee's work sample on a classification scale (e.g., below basic, basic, proficient, and advanced, with each category divided further into low, middle, and high). Then there is discussion by panelists of work samples with discrepant scores in order to share insights that might have been overlooked or missed by other panelists. Finally, there is re-rating of the set of examinee work samples. This procedure is repeated for all of the questions that comprise the assessment.

Work samples classified into the borderline categories are used to calculate performance standards. The average score of the work samples assigned by the panelists to the relevant borderline category become the recommended point estimate of that performance standard. For example, for setting a "proficient" standard, all of the work samples classified as "high," "basic," and "low" proficient are used in calculating the average score. This average score of the borderline examinees is used as the performance standard. The process is repeated for basic and advanced standards, and for all the questions in the assessment. Then, the total assessment standards are obtained by summing over the standards set (e.g., basic, proficient, advanced) on the individual questions.

The AJM holds promise for use with complex performance tasks. Pilot studies in three states indicate that the method is easy to use and that it results in cutoff scores that panelists feel are appropriate. However, other validity data are needed to examine the accuracy of these performance standards (Plake & Hambleton, 2001).

Guidelines for Practice

On the basis of the information above, as well as two reviews of the literature on cutoff

scores (Cascio et al., 1988; Truxillo, Donahue, & Sulzer, 1996), we suggest the following:

- Follow Standard 4.19 (AERA, APA, & NCME, 1999) regarding the description and documentation of the method used, the selection and training of judges, and an assessment of their variability. These recommendations are sound no matter which specific method of setting cutoff scores are used.
- Determine if it is necessary to set a cutoff score at all, as legal and professional guidelines do not demand their use in all situations.
- It is unrealistic to expect that there is a single "best" method of setting cutoff scores for all situations.
- Begin by identifying relative levels of proficiency on critical knowledge, skills, abilities, and other characteristics (e.g., using the analytical judgment method).
- If a cutoff score is to be used as an indicator of minimum proficiency, relating it to what is necessary in the educational environment, on the job, or other relevant context is essential. The Angoff method may be helpful in doing this.
- When using judgmental methods, sample a sufficient number of subject matter experts (SMEs). That number usually represents about a 10–20% sample of job incumbents and supervisors, representative of the race, gender, location, shift, and assignment composition of the entire group. However, the most important demographic variable in SME groups is experience (Landy & Vasey, 1991). Failure to include a broad cross-section of experience in a sample of SMEs could lead to distorted ratings.
- Consider errors of measurement and adverse impact when setting a cut score. Thus, if the performance of incumbents is used as a basis for setting a cutoff score that will be applied to a sample of applicants, it is reasonable to set the cutoff score one standard error of measurement below the mean score achieved by incumbents.
- Set cutoff scores high enough to ensure that minimum standards of performance are met. Either the Angoff or AJM procedures can help to determine what those minimum standards should be.

It is unrealistic to expect that there is a single "best" method of setting cutoff scores for all situations.

Cross-Validation

HR practitioners who develop and use tests are concerned with the prediction of behaviors and outcomes (e.g., job performance) based on some available information (e.g., pre-employment test scores). This prediction is often implemented by linking certain information (i.e., predictors) with the desired outcome (i.e., criterion), assuming a linear relationship between the predictors and the criterion (i.e., one best described by a straight line). These relationships are typically operationalized using ordinary least squares (OLS) regression, in which weights are assigned to the predictors so that the difference between observed criterion scores and predicted criterion scores is minimized (Appendix B includes additional technical information).

Although OLS regression is arguably the most commonly used prediction technique in practice, the assumption that regression weights obtained from one sample can be used with other samples with a similar level of predictive effectiveness is not true in most situations. In other words, the computation of regression weights is affected by idiosyncrasies of the sample on which they are computed and it capitalizes on chance factors so that prediction is optimized in the sample. Thus, when weights computed in one sample (i.e., experienced employees) are used with a second sample from the same population (i.e., newly hired employees), the multiple correlation coefficient is likely to be smaller. This phenomenon has been labeled *shrinkage* (Larson, 1931).

Of course, the CHRO in our opening vignette is not interested in predicting specific outcomes in one sample only. Thus, an important question is the extent to which weights derived from a sample cross-

validate (i.e., generalize). Cross-validity (i.e., ρ_c) refers to whether the weights derived from one sample can predict outcomes to the same degree in the population as a whole or in other samples drawn from the same population. If cross-validity is low, the use of assessment tools and prediction systems derived from one sample may not be appropriate in other samples from the same population.

Cross-validity (i.e., ρ_c) refers to whether the weights derived from one sample can predict outcomes to the same degree in the population as a whole or in other samples drawn from the same population.

Empirical and Statistical Strategies for Estimating Cross-Validity

The following two strategies are used to estimate cross-validity: (a) empirical and (b) statistical. The empirical strategy consists of fitting a regression model in a sample and using the resulting regression weights with a second, independent cross-validation sample. The multiple correlation coefficient obtained by applying the weights from the first (i.e., "derivation") sample to the second (i.e., "cross-validation") sample is used as an estimate of ρ_c. Alternatively, only one sample is used, but it is divided into two subsamples, thus creating a derivation and a cross-validation subsample. This is known as a single-sample strategy.

The statistical strategy consists of adjusting the sample-based, multiple correlation coefficient (R) by a function of sample size (N) and the number of predictors (k) (see Appendix B for technical details). An alternative statistical strategy to cross-validation is the jackknife method (Efron & Gong, 1983). The name *jackknife* was coined by Tukey (1958) to imply that the method is an all-purpose statistical tool. It consists of obtaining random subsamples with replacement from the full original sample. Thousands of such subsamples are generated and the validity coefficient is computed for each. Then, a distribution of validity coefficients is obtained and the mean validity coefficient across all the subsamples is computed. This mean validity coefficient is used as an estimate of cross-validity.

Although empirical (e.g., Browne, 2000; Mosier, 1951) and statistical (Ezekiel, 1930; Wherry, 1931) strategies for cross-validation have been available for more than half a century, recent research provides new insights into the advantages and disadvantages of each of the two approaches.

Comparison of Empirical and Statistical Strategies

Given the strategies available, several authors have compared empirical and statistical approaches to cross-validation (e.g., Lautenschlager, 1990; Raju, Bilgic, Edwards, & Fleer, 1997, 1999; Schmitt & Ployhart, 1999). This research adds to our knowledge base from the late 1970s (e.g., Drasgow, Dorans, & Tucker, 1979; Schmitt, Coyle, & Rauschenberger, 1977) as well as 1980s (Cotter & Raju, 1982; Mitchell & Klimoski, 1986; Murphy, 1983, 1984). Given the results of these investigations, we are now in a position to draw some conclusions regarding the estimation of cross-validity.

Empirical approaches. The trade-offs involved in the use of single-sample designs suggest that such designs are rarely justified (Murphy, 1983). First, splitting the sample causes a loss of degrees of freedom; subsequently, regression weights become unstable. Second, a single-sample approach accounts for random error, but not for systematic error, and this may lead researchers to overestimate cross-validity. For instance, assume that a sample used in a single-sample design, cross-validation effort is biased and not representative of the population. For example, suppose a sample of college-graduate accountants is used to validate a test of accounting aptitude that is intended for use with high-school students. If the sample is randomly divided to create derivation and cross-validation subsamples of college-graduate accountants, these two samples will share characteristics with each other, but will not share characteristics with the population of high-school students, or other nonbiased samples drawn from the same population. Thus, the regression weights obtained from the derivation sample may perform equally well in the cross-validation sample, and a researcher may

conclude that cross-validity is high. However, this result would not replicate in other nonbiased samples, and, therefore, the conclusion regarding cross-validity would be erroneous (Murphy, 1984).

The implementation of multiple-sample designs demands more effort and cost on the part of researchers. Consequently, they are implemented infrequently (Murphy, 1984). In addition, results show that multiple-sample designs seem to yield accurate estimates of cross-validity only when the validation sample is representative of the population (Claudy, 1978), and the derivation sample is large relative to *k*, the number of predictors.

Statistical approaches. Statistical approaches are more cost-effective to implement. Thus, if a formula-based approach to estimating cross-validity in the population is as accurate as the empirical approaches, then the formula-based approach would be the preferred strategy.

Several comparisons of cross-validity estimation procedures have been conducted over the last three decades (e.g., Browne, 2000; Cotter & Raju, 1982; Schmitt et al., 1977). The most comprehensive comparison to date is Raju et al. (1999), who investigated 11 cross-validity estimation procedures. The overall conclusion of this body of research is that Equation B2 (see Appendix B) provides accurate results as long as *N* (the sample size) is greater than 40.

Cross-Validation in Practice

In spite of the abundant body of research generated by quantitative psychologists, it seems that, overall, HR practitioners and social scientists interested in assessment have not paid much attention to cross-validation issues. For example, Mitchell (1985) reviewed 126 articles published in the *Journal of Applied Psychology, Organizational Behavior and Human Performance,* and the *Academy of Management Journal* between 1979 and 1983 and found that only seven (5.5%) attempted cross-validation. More recently, St. John and Roth (1999) reviewed articles published in the *Academy of Management Journal, Administrative Science Quarterly,* and the *Strategic Management Journal* between January 1990 and December 1995 and found that none of the articles reviewed reported empirical or formula-based cross-validation estimates. The only exception to the rule seems to be the area of empirical keying of noncognitive predictors such as biographical inventories (i.e., biodata). The process of selecting biodata items includes a built-in, empirical cross-validity procedure in which the final items are chosen based on their ability to discriminate high from low performers in the cross-validation sample based on weights obtained in the derivation sample (e.g., Mael & Hirsch, 1993).

Every *study involving assessment should include cross-validation estimates.*

Guidelines for Practice

- HR practitioners interested in assessment should pay greater attention to the issue of cross-validation. *Every* study involving assessment should include cross-validation estimates. Consumers of assessment tools should demand cross-validation information before deciding to use specific tests.
- Test users and developers should not confuse the often-reported "adjusted R^2" with the cross-validity coefficient. Be aware that the adjusted R^2 underestimates the amount of capitalization on chance and overestimates the proportion of variance explained in the criterion. The adjusted R^2 is only an intermediate step in computing the cross-validity coefficient.
- Logistical considerations, as well as the cost associated with the conduct of empirical cross-validation studies, can be quite demanding. There seem to be no clear advantages to using empirical designs in most situations, and results of empirical research indicate that a statistical approach is as accurate as an empirical approach in most situations. The use of the jackknife method is a

good statistical alternative, but it can be time-consuming and not easily accessible to most HR practitioners. Thus, we suggest that Equation B2 be used to estimate cross-validity (see Appendix B).

Scientists clearly have much to learn from practitioners about the process of implementation.

Concluding Remarks

Our discussion reveals that recent research-based advances regarding (a) validity generalization, (b) statistical significance testing, (c) criterion measures, (d) cutoff scores, and (e) cross-validation have five important implications, not only for the CHRO described in the opening vignette, but also for any HR practitioner who uses assessment procedures.

1. Validity generalization has helped disconfirm the hypothesis that validity varies from situation to situation, but it is as fallible as any other data-analytic procedure, it should not be used indiscriminately, and it should not serve as the sole method for defending the use of an assessment procedure.
2. HR practitioners who want to learn whether a conclusion based on a sample regarding test validity generalizes to the relevant population should consider reporting statistical power and confidence intervals, in addition to significance levels.
3. In developing and using criterion measures, practitioners should recognize that criteria are dynamic and multidimensional and should specify whether they are trying to predict typical or maximum performance.
4. If HR practitioners determine that it is necessary to establish a cutoff score as an indicator of minimum proficiency, the cut score should be based on a valid assessment procedure and reflect minimum qualification standards.
5. *Every* predictive study should include a cross-validation estimate. Recognize that adjusted R^2 is only an intermediate step in this process.

As we noted at the outset, each of the areas we reviewed is controversial. Even experts disagree about appropriate ways to address them, although there is general agreement about the guidelines we present.

While the guidelines presented here are sound technically, it is important to stress that technically meritorious practices are sometimes not adopted for at least three reasons (Johns, 1993). One, managers frame HR practices as matters of administrative style rather than as technical innovations. Two, industrial/organizational psychologists often justify HR practices from a technical perspective only, ignoring important social and contextual influences that affect the adoption of innovations. Three, crises, organizational politics, competing sources of innovation, government regulation, and institutional factors often overshadow technical merit. We hope our article presents technical information with sufficient clarity that HR practitioners will see the applicability and usefulness of this information.

The process of implementing recommendations relevant to the five areas covered here is beyond the scope of this article. Clearly, there is a gap between scientists and practitioners, although opinions differ about whether it is growing (Hulin, 2001) or shrinking (Latham, 2001). As Muchinsky (2004) has noted, scientists, for the most part, are relatively unconcerned about how their theories, principles, and methods are put into practice outside of academic study. For the most part, practitioners are deeply concerned with matters of implementation because what they do occurs in arenas not created primarily for scientific study. A key criterion for practitioners, therefore, is "organizational acceptability." Scientists clearly have much to learn from practitioners about the process of implementation. To the extent that both parties are willing, even eager, to learn from each other, there is hope that the gap that separates them might someday be difficult to discern.

The research reported in this article was supported, in part, by a Summer Grant from The Business School of the University of Colorado at Denver and Health Sciences Center.

Wayne F. Cascio, PhD, is the US Bank Term Professor of Management at the University of Colorado at Denver and Health Sciences Center. He is past chair of the HR Division of the Academy of Management and past president of the Society for Industrial and Organizational Psychology. An elected fellow of the Academy of Management, he received the Distinguished Career award from the Academy's HR Division in 2000, and an honorary doctorate from the University of Geneva (Switzerland) in 2004. He serves on the boards of directors of CPP, Inc., the Society for Human Resource Management Foundation, and the Academy of Management.

Herman Aguinis, PhD, is the Mehalchin Term Professor of Management in the Business School at the University of Colorado at Denver and Health Sciences Center. He is the author of *Regression Analysis for Categorical Moderators* (2004, Guilford), coauthor (with Wayne Cascio) of *Applied Psychology in Human Resource Management* (2005, Prentice Hall), and editor of *Test Score Banding in Human Resource Selection* (2004, Praeger). He has published more than 50 articles in the *Academy of Management Journal,* the *Academy of Management Review,* the *Journal of Applied Psychology, Personnel Psychology, Organizational Behavior and Human Decision Processes,* and other refereed journals. His research interests include staffing, social power and influence in organizations, and research methods and analysis in management and related fields.

NOTE

1. Throughout this article, the technical terms that are underlined are defined in Appendix A.

REFERENCES

Aguilera v. Cook County Police and Corrections Merit Board. 760 F.2d 844 (7th Circuit, 1985).

Aguinis, H. (2001). Estimation of sampling variance of correlations in meta-analysis. Personnel Psychology, 54, 569–590.

Aguinis, H. (2004). Regression analysis for categorical moderators. New York: Guilford.

Aguinis, H., & Whitehead, R. (1997). Sampling variance in the correlation coefficient under indirect range restriction: Implications for validity generalization. Journal of Applied Psychology, 82, 528–538.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508–600). Washington, DC: American Council on Education.

Bernard v. Gulf Oil Corp. 890 F.2d 735 (5th Circuit, 1989).

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), Personnel selection in organizations (pp. 71–98). San Francisco: Jossey-Bass.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. Human Performance, 10, 99–109.

Browne, M. W. (1975). Predictive validity of a linear regression equation. British Journal of Mathematical and Statistical Psychology, 28, 79-87.

Browne, M. W. (2000). Cross-validation methods. Journal of Mathematical Psychology, 44, 108–132.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), Personnel psychology in organizations (pp. 35–70). San Francisco: Jossey-Bass.

Cascio, W. F. (2000). Costing human resources: The financial impact of behavior in organizations (4th ed.). Cincinnati, OH: South-Western College Publishing.

Cascio, W. F., & Aguinis, H. (2001). The federal Uniform Guidelines on Employee Selection Procedures (1978): An update on selected

issues. Review of Public Personnel Administration, 21, 200–218.

Cascio, W. F., & Aguinis, H. (2005). Applied psychology in human resource management (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. Personnel Psychology, 41, 1–24.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. Applied Psychological Measurement, 2, 595–607.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coleman, V. I., & Borman, W. C. (2000). Investigating the underlying structure of the citizenship performance domain. Human Resource Management Review, 10, 25–44.

Cotter, K. L., & Raju, N. S. (1982). An evaluation of formula-based population squared cross-validity estimates in prediction. Educational and Psychological Measurement, 40, 101–112.

Deadrick, D. L. & Madigan, R. M. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. Personnel Psychology, 43, 717–744.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. Applied Psychological Measurement, 3, 171–192.

DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. Journal of Applied Psychology, 78, 205–211.

EEOC v. Atlas Paper Box Co. 868 F.2d at 1487 (6th Circuit, 1989).

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37, 36–48.

Ezekiel, M. (1930). Methods of correlational analysis. New York: Wiley.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.

Hofmann, D. A., Jacobs, R., & Baratta, J. E. (1993). Dynamic criteria and the measurement of change. Journal of Applied Psychology, 78, 194–204.

Hulin, C. L. (2001). Applied psychology and science: Differences between research and practice. Applied Psychology: An International Review, 50, 225–234.

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. Psychological Bulletin, 107, 328–340.

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

Johns, G. (1993). Constraints on the adoption of psychology-based personnel practices: Lessons from organizational innovation. Personnel Psychology, 46, 569–592.

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. Journal of Applied Psychology, 86, 984–996.

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. Psychological Bulletin, 127, 673–697.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. Journal of the American Statistical Association, 44, 1372–1381.

Landy, F. J. (2003). Validity generalization: Then and now. In K. R. Murphy (Ed.), Validity generalization: A critical review (pp. 155–195). Mahwah, NJ: Lawrence Erlbaum Associates.

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. Personnel Psychology, 44, 27–50.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. Journal of Educational Psychology, 22, 45–55.

Latham, G. P. (2001). The reciprocal transfer of learning from journals to practice. Applied Psychology: An International Review, 50, 201–211.

Lautenschlager, G. L. (1990). Sources of imprecision in formula cross-validated multiple correlations. Journal of Applied Psychology, 75, 460–462.

Mael, F. A., & Hirsch, A. C. (1993). Rainforest empiricism and quasi-rationality: Two approaches to objective biodata. Personnel Psychology, 46, 719–738.

Markus, K. A. (2001). The converse inequality argument against tests of statistical significance. Psychological Methods, 6, 147–160.

Mitchell, T. W. (1985). An evaluation of the validity of correlational research conducted in organiza-

tions. Academy of Management Review, 10, 192–205.

Mitchell, T. W., & Klimoski, R. J. (1986). Estimating the validity of cross-validity estimation. Journal of Applied Psychology, 71, 311–317.

Mosier, C. I. (1951). Symposium: The need and means of cross-validation. Educational and Psychological Measurement, 11, 5–11.

Muchinsky, P. M. (2004). When the psychometrics of test development meets organizational realities: A conceptual framework for organizational change, examples, and recommendations. Personnel Psychology, 57, 175–209.

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single sample designs. Personnel Psychology, 36, 111–118.

Murphy, K. R. (1984). Cost-benefit considerations in choosing among cross-validation methods. Personnel Psychology, 37, 15–22.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods, 5, 241–301.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. Human Performance, 10, 85–97.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), Setting performance standards: Concepts methods, and perspectives (pp. 283–312). Mahwah, NJ: Lawrence Erlbaum.

Raju, N. S., Anselmi, T. V., Goodman, J. S., & Thomas, A. (1998). The effect of correlated artifacts and true validity on the accuracy of parameter estimation in validity generalization. Personnel Psychology, 51, 453–465.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. Applied Psychological Measurement, 21, 291–305.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weight procedures. Applied Psychological Measurement, 23, 99–115.

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. Human Resource Management, 41, 149–174.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73, 482–486.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115–129.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529–540.

Schmidt, F. L, & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124, 262–274.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419–434.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 84, 751–758.

Schmitt, N., & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. Journal of Applied Psychology, 84, 50–57.

Society for Industrial and Organizational Psychology. (SIOP) (2003). Principles for the validation and use of personnel selection procedures (4th ed.). Bowling Green, OH: Author.

Steele-Johnson, D., Beauregard, R. S., Hoover, P. B., & Schmidt, A. M. (2000). Goal orientation and task demand effects on motivation, affect, and performance. Journal of Applied Psychology, 85, 724–738.

Steele-Johnson, D., Osburn, H. G., & Pieper, K. F. (2000). A review and extension of current models of dynamic criteria. International Journal of Selection and Assessment, 8, 110–136.

St. John, C. H., & Roth, P. L. (1999). The impact of cross-validation adjustments on estimates of effect size in business policy and strategy research. Organizational Research Methods, 2, 157–174.

Task Force on Statistical Inference. (2000). Narrow and shallow. American Psychologist, 55, 965.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. Human Performance, 9, 275–295.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. Psychological Methods, 6, 371–386.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. Annals of Mathematical Statistics, 29, 614.

Uniform Guidelines on Employee Selection Procedures. (1978). Federal Register, 43, 38290–38315.

Wainer, H. (1999). One cheer for null hypothesis significance testing. Psychological Methods, 4, 212–213.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of multiple correlation. Annals of Mathematics and Statistics, 2, 440–457.

Appendix A

Definitions of Key Technical Terms

Adverse impact—refers to a substantially different rate of selection in hiring, promotion, or other employment decision that works to the disadvantage of members of a race, gender, or ethnic group.

Correlation coefficient—a measure of the overall strength of relationship between two variables. It varies from –1 to +1. A value of zero indicates no relationship.

Criterion measures—outcomes of interest (e.g., measures of job performance).

Degrees of freedom—in estimating a population parameter (e.g., a mean), the number of observations in a distribution that are free to vary.

Effect size—the degree of departure from the null hypothesis that there is no relationship between two variables (e.g., a correlation coefficient or a measure expressed in standard deviation units).

Mean—the average of a set of values.

Null hypothesis—the hypothesis that there is no effect or relationship (e.g., there is no relationship between scores on an assessment and scores on a measure of performance).

Population—a complete collection of observations, one that contains every data point of a certain grouping.

Predictors—assessment tools (e.g., tests, interviews) used to forecast performance.

Random error—any deviation from a true population value that results from chance fluctuation.

Range restriction—inclusion of less than 100 percent of the full range of variables (e.g., use of only five points on a seven-point rating scale).

Sample—a subset of a larger population.

Sampling error—the difference between sample and population values that is due to the particular units selected for observation.

Significance testing—in the context of classical hypothesis testing, if an observed value exceeds or falls below an expected value by some amount, such that the deviation is unlikely to have occurred by chance alone (e.g., at a probability of 1 in 20), the deviation is said to be "statistically significant."

Standard deviation (SD)—a measure of variability around an average value. In a normal or bell-shaped distribution, three SDs above and below the mean include more than 99 percent of all observations. Two SDs include more than 95 percent of all observations.

Standard error of measurement—the standard deviation of an applicant's score distribution if she or he were to take the assessment repeatedly with no new learning taking place between administrations and no memory of prior questions.

Statistical power—the likelihood of correctly concluding that an effect exists, if it is indeed present (e.g., that a relationship between test scores and performance exists).

Subject matter experts (SMEs)—individuals chosen for their expertise in a particular area to provide input to a management activity (e.g., job analysis, development of assessment tools).

Systematic error—a deviation of the same amount or degree from a true population value (e.g., a watch that is always five minutes fast).

Validity coefficient—the value of the correlation coefficient that reflects the overall strength of relationship between predictor and criterion scores.

Appendix B

Technical Information on Cross-Validation

Ordinary Least Squares Regression

Assuming a prediction situation that includes two predictors (P_1 and P_2), the OLS regression equation is the following:

$$\hat{Y} = a + b_1P_1 + b_2P_2 + \varepsilon \qquad \text{(B1)}$$

where $\hat{Y}$ is the predicted value for Y, a is the estimate of the intercept, b_1 is the regression weight for P_1, b_2 is the regression weight for P_2, and ε is error (Aguinis, 2004). Once weights are computed for each of the predictors based on a sample, test developers hope, and often assume, that the same weights will provide an equally optimal prediction system for other samples drawn from the same population. That is, it is generally assumed that a regression equation linking pre-employment test scores with supervisory job performance ratings for current employees would be just as effective in predicting job performance for the newly hired employees.

Statistical Cross-Validation

Numerous formulas are available to implement the statistical strategy (Raju et al., 1997). The most commonly implemented formula to estimate cross-validity (i.e., ρ_c) is the following (Browne, 1975):

$$\rho_c^2 = \frac{(N-k-3)\rho^4 + \rho^2}{(N-2k-2)\rho^2 + \rho} \qquad \text{(B2)}$$

where ρ is the population multiple correlation, N is the sample size, and k is the number of predictors. The squared multiple correlation in the population, ρ^2, can be computed as follows (Ezekiel, 1930, and usually attributed to Wherry, 1931):

$$\rho^2 = 1 - \frac{(N-1)}{(N-k-1)}(1-R^2) \qquad \text{(B3)}$$

Note that Equation B3 is what most computer outputs label "adjusted R^2," but it is only an *intermediate step* in computing cross-validity (i.e., Equation B2). Equation B3 does not directly address the capitalization on chance in the sample used and only addresses the issue of shrinkage partially by adjusting the multiple correlation coefficient based on sample size and the number of predictors in the regression model (St. John & Roth, 1999). The obtained "adjusted R^2" does not address the issue of prediction optimization due to sample idiosyncrasies and, therefore, underestimates the shrinkage. The use of Equation B3 *in combination* with Equation B2 addresses this issue.