PERSONNEL PSYCHOLOGY
2005, 58, 1009–1037

A META-ANALYSIS OF WORK SAMPLE TEST VALIDITY: UPDATING AND INTEGRATING SOME CLASSIC LITERATURE

PHILIP L. ROTH
Department of Management
Clemson University

PHILIP BOBKO
Department of Management
Gettysburg College

LYNN A. McFARLAND
Department of Psychology
Clemson University

Work sample tests have been used in applied psychology for decades as important predictors of job performance, and they have been suggested to be among the most valid predictors of job performance. As we examined classic work sample literature, we found the narrative review by Asher and Sciarrino (1974) to be plagued by many methodological problems. Further, it is possible that data used in this study may have influenced the results (e.g., $r = .54$) reported by Hunter and Hunter in their seminal work in 1984. After integrating all of the relevant data, we found an observed mean correlation between work sample tests and measures of job performance of .26. This value increased to .33 when measures of job performance (e.g., supervisory ratings) were corrected for attenuation. Our results suggest that the level of the validity for work sample tests may not be as large as previously thought (i.e., approximately one third less than previously thought). Further, our work also summarizes the relationship of work sample exams to measures of general cognitive ability. We found that work sample tests were associated with an observed correlation of .32 with tests of general cognitive ability.

Work sample tests are generally thought to have a number of very desirable attributes. They are believed to be among the most valid predictors of job performance by researchers (Hunter & Hunter, 1984; Reilly & Warech, 1993) and managers (see Terpstra, Kethley, & Foley, 2000). They are also believed by many researchers to have lower levels of standardized ethnic group differences (Cascio, 2003; Schmitt & Mills, 2001) and

We thank many researchers for their help in this work. We thank Frank Schmidt, Deniz Ones, Vish C. Viswesvaran, Mike McDaniel, and Paul Sackett for their efforts to identify the studies used by Hunter and Hunter in their seminal work. We also thank Heinz Schuler for his help in locating unpublished studies in this area.

Correspondence and requests for reprints should be addressed to Philip Roth, Clemson University, Department of Management, Clemson, SC 29634; rothp@clemson.edu.

adverse impact (Callinan & Robertson, 2000) than other predictors of job performance such as cognitive ability tests. Finally, work sample tests are thought to be viewed positively by job applicants (Hattrup & Schmitt, 1990; Hausknecht, Day, & Thomas, 2004; Vance, Coovert, MacCallum, & Hedge, 1989).

Nevertheless, there is also an important set of limitations regarding the field's evidence of the validity of work sample tests. First, major meta-analyses in this area occurred over 20 years ago (e.g., Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984) despite the fact that substantial amounts of new data are available. Second, there are a number of confusing and confounding issues in prior meta-analyses and narrative reviews. In one instance, there is no record of the studies that were used in the meta-analysis (Hunter & Hunter, 1984), and another meta-analysis—likely due to a more encompassing purpose—only covered two journals and analyzed only seven studies in which the dependent variable was job performance (Schmitt et al., 1984). An important early narrative review (i.e., Asher & Sciarrino, 1974) identified a wide variety of tests as work samples (e.g., job knowledge, situational judgment, etc.) and is plagued with a number of methodological problems. Third, there has been comparatively little attention to moderators in work sample meta-analyses. All told, there is an important opportunity to clarify and update the previous summaries of studies on work sample validity.

The purpose of this article is to meta-analyze the validity of work sample tests for predicting job performance (i.e., supervisory ratings and objective measures of job performance). An important part of this work was reviewing all of the previous literature cited in major reviews and meta-analyses to make sure that (a) the predictors were appropriate to analyze as "work samples" and (b) the data were free of methodological problems that might bias validity estimates. We also examined a number of moderators of validity. Finally, we analyzed data describing the relationships between work samples and several other predictors of job performance. In this way, we hope to update the understanding of this well-regarded predictor of job performance.

Defining the Work Sample Test

We examined a number of sources as we considered the definition of a work sample test (e.g., Gatewood & Field, 2001; Guion, 1998; Ployhart, Schneider, & Schmitt, in press) and noted their substantial degree of commonality. We adopt the definition of Ployhart et al. (in press) who state "a work sample test is a test in which the applicant performs a selected set of actual tasks that are physically and/or psychologically similar to those performed on the job." Ployhart et al. also note the importance of structure to work sample tests by stating that "procedures are standardized

and scoring systems are worked out with the aid of experts in the occupation in question." Guion (1998) emphasizes that one of the defining characteristics of a work sample test is a relatively high level of fidelity (i.e., low level of abstraction) with the job in question. Heneman and Judge (2003) also provide an important distinction in defining work sample tests and distinguishing them from another form of testing. They use the term "performance test" to refer to a situation in which applicants actually do the job (e.g., internships, job tryouts, and probationary periods). We concur with this distinction and did not use any performance tests in our analyses.

Current Understanding of the Validity of Work Sample Tests

Our current understanding of work sample validity is heavily influenced by the ground-breaking work of Hunter and Hunter (1984). This oft-cited, pioneering article reported that the validity of work sample tests for predicting supervisory ratings was .54. This estimate of validity, corrected for the unreliability of supervisory ratings, has continued to be cited as one of the most definitive estimates to date (e.g., see Schmidt & Hunter, 1998). This value is particularly interesting because Hunter and Hunter report that, for experienced workers, the validity of work sample exams is slightly higher than the validity of .51 for cognitive ability tests. Consequently, work sample tests are thought by many applied psychologists to be among the most valid predictors of job performance.

The value of .54 is also interesting because it is not possible to determine the studies that went into this estimate. One possible reason for the lack of a list of studies in the work sample analysis is that conventions for reporting primary studies were not well established at this time in the development of meta-analysis.

Hunter and Hunter (1984) also meta-analytically reanalyzed the data of a prominent review article by Asher and Sciarrino (1974). In this earlier article, researchers defined work sample tests as "a miniature replica of the criterion task" (p. 519). Analyses for work sample tests were broken down into the categories of psychomotor tests and verbal tests. The Hunters' meta-analytic reanalysis resulted in validities of .62 for motor tests predicting job performance and .45 for verbal tests predicting performance (see Hunter & Hunter, 1984, Table 6, p. 84) and these coefficients appeared to have been corrected for unreliability in supervisory performance ratings.

It is important to note that Hunter and Hunter (1983, 1984) did not explicitly endorse these tests as being work samples in their analysis. In fact, Hunter and Hunter cautioned readers that many of the tests in Asher and Sciarrino (1974) were likely to be viewed as job knowledge tests (see Hunter & Hunter, 1984, p. 84). Nevertheless, Asher and Sciarrino's work is still prominent in two ways. First, the Hunters explicitly re-analyzed the studies of Asher and Sciarrino as stated in the previous paragraph (see

also the thorough work of Salgado, Viswesvaran, and Ones (2001), for a current example of the prominence of Asher & Sciarrino). Second, it is also possible that Hunter and Hunter's meta-analytic results may have included the studies reviewed in Asher and Sciarrino.

Hunter also performed another apparently smaller-scale meta-analysis of work sample tests in an important book chapter (Hunter, 1983a) that involved nonmilitary studies and showed a work sample–supervisory rating mean correlation of .42 ($K = 7$, $N = 1{,}790$) when corrections for criterion unreliability were made. In a similar fashion, Hunter analyzed military studies and reported a mean corrected correlation of .27 ($K = 4$, $N = 1474$).

Around the same time, another team of researchers also estimated the validity of work sample tests (Schmitt et al., 1984). These researchers limited their primary studies to two leading journals—the *Journal of Applied Psychology* and *Personnel Psychology*—for the years between 1964 through 1984 so as to be able to compare a wide variety of predictors of job performance. They found an uncorrected validity of work samples of .378 ($K = 18$, $N = 3{,}512$) for predicting a variety of criteria (e.g., ratings of job performance, achievement/grades, wages, and work samples). When the criteria were narrowed to job performance ratings, work sample uncorrected validity was estimated to be .32 ($K = 7$, $N = 384$). Interestingly, Schmitt et al. also report that a variety of predictors were able to predict work samples used as a criterion. This indicates that work sample tests can be conceptually viewed as either predictors of job performance or as a criterion of job performance. We will return to this issue later. Schmitt et al.'s overall analyses were updated by Russell and Dean (1994) by adding data from the years of 1984–1992. The observed validity of work sample tests was estimated to be .373 ($K = 20$, $N = 3894$) for a variety of jobs after Russell and Dean added two new studies with a total of approximately 400 participants.

Limitations of Previous Work

The previously noted studies have made important contributions to the understanding of work sample test validity. Nevertheless, there are several major limitations of these studies that suggest an update and re-analysis is needed.

First, there has not been a major, field-wide review of the validity of work sample tests since the work of Hunter and Hunter (1984) 20 years ago (recall Russell & Dean's work applied to only two journals). This passage of time means that additional studies on work sample validity are available. In fact, there have been many studies involving thousands of participants that have been conducted over the past 2 decades. Further,

technological developments in the ability to search relevant databases can also help researchers identify and find these primary studies.

Second, there was a very loose application of the definition of work sample exams in the work of Asher and Sciarrino (1974). Normally, one would look at this review as somewhat historic given its age of 30+ years. Nevertheless, the potential influence it may have had on our current understanding of work sample exams (via the possible link with Hunter & Hunter, 1984) is quite important. Combining the methodological limitations of Asher and Sciarrino's work with the limited scope (i.e., two journals) of other researchers' work leads one to question how comprehensive and appropriate our estimates of work sample validity are at the present time.

By today's standards, it appears that Asher and Sciarinno (1974) used an extremely wide variety of test types in their analysis. We provide a few examples below, but others appear in their paper.

(1) Asher and Sciarinno classified standardized job knowledge tests as work samples. For example, farm knowledge was assessed to facilitate vocational assessment and counseling for prisoners (Grigg, 1948), knowledge of insurance was assessed with a standardized multiple-choice paper-and-pencil measure to assess salespeople (Baier & Dugan, 1956), and common facts of law were assessed via true/false questions to test entering law students (Adams, 1948). Hunter and Hunter (1984) noted this limitation as well (p. 84).
(2) Asher and Sciarrino classified what we would now call situational judgment tests as work samples. For example, Forehand and Guetzkow (1961) administered a multiple- choice administrative judgment test to managers in the federal government, Knauft (1949) used a multiple-choice test of judgment in material problems to predict the performance of bake shop managers, and Mandell (1947) used a multiple-choice test of administrative judgment.
(3) In other cases, researchers used paper-and-pencil measures of cognitively related abilities that did not appear to be directly representative of the tasks of the job. For example, McNamara and Hughes (1961) appeared to use standardized tests of paper-and-pencil tests of mathematical and reasoning abilities to predict computer programming performance.

The key issue here is that these examples, and many more studies in Asher and Sciarrino (1974), do not appear to fit their own definition of a work sample as a miniature replica of the criterion. Further, many of the tests in Asher and Sciarrino do not fit our definition as "tasks that are physically and/or psychologically similar to those performed on the job."

It is also worth reiterating that the re-analysis of Asher and Sciarrino's work continues to be cited in modern literature reviews (Salgado et al., 2001).

Third, many of the coefficients used by Asher and Sciarrino (1974) were subject to range enhancement (Hunter & Schmidt, 1990, 2004). Range enhancement has also been described by Bobko (2001) as "reverse range restriction." Just as it sounds, reverse range restriction occurs when the range of values on a variable or variables is artificially increased. For example, reverse range restriction occurs when only the highest third of individuals and the lowest third of individuals are entered into any analysis. Deletion of the middle third of individuals increases the variance of values (as values close to the mean are not used in analysis) and correlations are too large. Inclusion of such range-enhanced statistics in a meta-analysis could bias the final estimate of ρ upward. Examples of range enhancement include Abt (1949), Blum (1943), Drewes (1961), Glaser, Schwarz, and Flanagan (1958), Knauft (1949), and Poruben (1950). Interestingly, Blum notes the issue of upward bias in his paper so that at least some researchers were aware of this issue in the literature that Asher and Sciarrino reviewed.

Finally, other coefficients in Asher and Sciarrino (1974) appear to be "contaminated." That is, the individual(s) making performance judgments also either had knowledge of the work sample exam scores or gave the work sample exam. Examples include Bender and Loveless (1958) and West and Bolanovich (1963). Similar problems also apply to some later primary studies not cited in Asher and Sciarrino (e.g., Robertson & Mindel, 1980). For those studies included in Asher and Sciarrino, this could also upwardly bias existing estimates of validity.

To summarize, there are important reasons to conduct a new meta-analysis on work sample exams. The work of Asher and Sciarrino (1974) is limited by the previously noted conceptual and methodological problems. It is difficult to know the source of data from the work of Hunter and Hunter (1984). A third meta-analysis (i.e., Schmitt et al., 1984) examined only two journals and used a small sample size for the dependent variable of job performance.

One other reason for our analysis involves the fact that there are relatively little data summarizing how various predictors of job performance relate to each other (Bobko, Roth, & Potosky, 1999). That is, how do various predictors of job performance intercorrelate? Researchers have called for such research for decades (Hunter & Hunter, 1984), but there are still relatively little data in this area (see also Schmidt & Hunter, 1998). Although not our primary focus, we also coded available correlations between work sample tests and three fairly common predictors of job performance: interviews, situational judgment tests, and measures of general cognitive ability.

Moderators of Validity

There are also a number of potential moderators of work sample validity.

Applicant versus incumbent. First, we examined if coefficients came from studies that were based on applicants or incumbents. Studies on applicants were predictive in nature such that the work sample exam was administered at one point in time (e.g., before hiring) and the measure of job performance was administered after hiring. We believe this moderator is important given the well-known effect that range restriction can have on correlations (when analyzing incumbents) and the opportunity that predictive studies allow to help correct such influences (Hunter & Schmidt, 1990, 2004). Meta-analyses on other variables bear out this thinking. For example, measures of personality from concurrent studies are higher than coefficients from predictive studies by an average of .07 (Hough, 1998). For situational judgment tests, researchers also found that the estimate of corrected validity from predictive studies was .18 ($K = 6$, $N = 346$) as opposed to .35 ($K = 96$, $N = 10{,}294$) for incumbent studies (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Finally, even Schmitt et al.'s (1984) analysis of many types of exams found that concurrent uncorrected validities were .34 ($K = 153$, $N = 17{,}838$) compared to .30 for predictive studies ($K = 99$, $N = 90{,}552$) to .26 for predictive studies using the test for selection ($K = 114$, $N = 124{,}960$).

Objective versus subjective measure of job performance. The issue of objective versus subjective measures of job performance has received substantial attention. The performance appraisal literature suggests that objective and subjective measures are not interchangeable (Murphy & Cleveland, 1991). Therefore, one might not expect substantial correlations between objective and subjective indicators of job performance. An earlier meta-analysis suggested that the results of objective and subjective measures of job performance had a mean observed correlation of .17 ($K = 23$, $N = 3{,}178$) that increased to .27 when corrected for attenuation (Heneman, 1986). A more recent meta-analysis found the observed correlation to be .32 ($K = 50$, $N = 8{,}341$) and the correlation, when corrected for unreliability on both variables, is .39 (Bommer, Johnson, & Rich, 1995).

In contrast, meta-analyses that focus on understanding the validity of various predictors of job performance often report that the nature of the criterion does not necessarily moderate the results. For example, there does not appear to be a moderating effect for type of criterion for either integrity tests (Ones, Viswesvaran, & Schmidt, 1993) or age (McEvoy & Cascio, 1989) as predictors of job performance. In sum, there appear to be conflicting views and more analysis may be needed.

Criterion versus predictor conceptualization. Work samples are somewhat unique because they can be considered either a predictor of job performance or a measure of job performance. In some studies, researchers conceptualized the work sample tests as predictors of ratings of job performance (Campion, 1972; Schmitt & Mills, 2001). In other studies, researchers conceptualized the work sample test as one of several criteria (e.g., Hunter, 1986; Nathan & Alexander, 1988; Schmidt, Hunter, Outerbridge, & Goff, 1988).

Previous research has not maintained or investigated this distinction, as most researchers have implicitly used all of the meta-analytic data available and conceptualized the work sample tests as predictors of job performance (e.g., Salgado et al., 2001; Schmidt & Hunter, 1998). We were interested to see if there was a moderating effect such that separating the data using the predictor and criteria conceptualization might result in different mean correlations between work samples and performance measures.

Military versus nonmilitary samples. We also coded coefficients as to whether they came from a military or nonmilitary sample. We pursued this coding in order to incorporate some of the distinctions of Hunter (1983a) in which he found marked differences between military and nonmilitary corrected correlations between work samples and supervisory ratings—mean *r*s of .27 and .42, respectively.

Job complexity. We also coded complexity of the job. Theoretically, we define job complexity to represent the information processing requirements of the job (rather than other components or requirements of the job). All other things being equal, more complex jobs might be reflected in more complex work sample exams and require higher levels of information processing. Noting the high validity of cognitive ability for predicting job performance (e.g., Schmidt & Hunter, 1998), it is possible that performance on work samples based on higher complexity jobs might have a stronger relationship to job performance.

Publication date. We coded the date each study was published. A reviewer suggested we use the publication date such that studies done up to 1982 could be contrasted with studies done after 1982. Such an analysis might suggest that the addition of new data was an important issue in estimating the validity of work sample tests.

Method

Literature Search

We searched for articles and manuscripts in several ways. First, we searched the electronic databases contained in PsycINFO, General Business File, Ingenta, and Expanded Academic ASAP. Second, we checked

all of the references that might be related to our work in previous reviews of work sample validity (Asher & Sciarrino, 1974; Hunter, 1983a; Russell & Dean, 1994; Schmitt et al., 1984), more general predictor review articles (Reilly & Warech, 1993) and other related work (Bommer et al., 1995; Roth, Huffcutt, & Bobko, 2003; Schmitt, Clause, & Pulakos, 1996). Third, we wrote to a number of researchers known for work in this area to assemble a list of the possible articles and technical reports that were available.

Inclusion Criteria

The issue of what criteria to use for inclusion was an important topic, given the conceptual and methodological issues we noted above. We used eight criteria for inclusion to provide our work with the opportunity to make estimates of work sample validity for predicting job performance that were as accurate as possible.

First, the dependent variable for all coefficients/studies was job performance. Studies measuring educational attainment or academic achievement were not included (e.g., Adams, 1943; Kazmier, 1962) as these were not the criteria that are typically used to validate personnel selection devices in the workplace.

Second, studies had to report data from participants who were either actual job applicants or incumbents employed in the job in question. This led us to exclude studies that used a work sample test within a laboratory experiment as either the performance task (e.g., Mount, Muchinsky, & Hanser, 1977; Pritchard, Hollenbeck, & DeLeo, 1980) or as a taped stimulus (e.g., Highhouse & Gallo, 1997). In addition, we excluded studies that focused on training the chronically underemployed or mentally ill (e.g., Franze & Ferrari, 2002), rehabilitating individuals with substance abuse problems such as drug use (e.g., Perone, DeWaard, & Baron, 1979), physical therapy rehabilitation (e.g., Callahan, 1993), and prison inmates (e.g., Grigg, 1948).

Third, studies had to report data from work samples that fit our earlier definition of work samples. Hence, we did not include situational judgment tests that were either recent (e.g., Lyons, Bayless, & Park, 2001) or from Asher and Sciarrino (1974; see examples above). Further, we did not include data from low-fidelity telephone simulations or assessment centers.

Our third criterion for inclusion also disqualified coefficients from studies in which subjects responded to questions by telling what they would do in actual work situations (e.g., some coefficients from Hedge, Teachout, & Laue, 1990). The approach of describing what one would do in a given work-related situation appears to be similar to the use of a situational

interview (Latham, Saari, Campion, & Purcell, 1980), and inclusion of this material would confound meta-analyses of the interviewing literature and the work sample literature.

We also excluded two classes of predictors previously used in analyses of work sample tests. As noted earlier, we did not include paper-and-pencil tests of job knowledge. It is also important to point out that we did not include coefficients from studies using leaderless group discussions if there was little or no documentation that this logically related to behaviors on the job (e.g., studies in Asher & Sciarrino, 1974; such as Bass 1954; Bass & Coates, 1954; Handyside & Duncan, 1954).

Fourth, studies had to report correlations based on "uncontaminated" data. That is, the individual who rated job performance could not be the same individual rating or supervising the work sample test.

Fifth, studies had to provide correlations that were not subject to range enhancement (Hunter & Schmidt, 1990, 2004) as noted above.

Sixth, studies had to provide independent correlations. Substantial care was taken to avoid coding and inappropriately analyzing "dependent" data in which results from multiple articles were reported based on the same samples (e.g., multiple military work samples analyzed by Hedge et al., 1990; Harville, 1996). The only situations in which we used two different coefficients from the same sample were when a study reported both objective and subjective measures of job performance. In these cases (specifically, three cases), both coefficients were only used in the moderator analysis of objective versus subjective performance measures. In all other analyses, the supervisory ratings were used from these articles.

Seventh, studies had to provide zero-order correlation coefficients or sufficient information to compute zero-order correlation coefficients. In some cases, correlations that were "partialled" for other variables (e.g., tenure or experience) were excluded from analysis (e.g., some coefficients from Scott, 1985).

Eighth, data in studies had to allow us to extricate the work sample test from other predictors in a battery. There were cases in which this was not possible because only composite validities of several predictors were reported (e.g., Davis, 1947; Gleason, 1957), and such data were not included in our analyses due to the extraneous variance of other predictors in a composite with work samples.

Coding Moderators

We coded several moderators. We dichotomously coded whether the participants in a study were applicants versus incumbents. We also dichotomously coded whether the measure of job performance was objective or subjective, whether the work sample was conceptualized as a predictor

versus a criterion, and whether the sample was from a military or nonmilitary organization. We coded job complexity based on the work of Hunter (1983b) that identified five levels of complexity depending upon the information processing requirements of the job. The five levels we coded were low (e.g., unskilled jobs such as receptionist), low medium (e.g., semi-skilled jobs such as truck driver), medium (e.g., skilled crafts such as electrician and first-line supervisors), medium high (e.g., computer trouble shooter), and high (e.g., scientists). We coded overall job performance as our dependent variable (e.g., overall supervisory ratings, summed ratings, or unit-weighted composites of multiple dimensions of performance). The first and third authors independently coded each of the studies. When there was disagreement, discrepancies were resolved through consensus.

Analyses

We meta-analyzed the uncorrected correlation coefficients from the studies in the reference list noted with an asterisk. That is, we recorded the observed *r* rather than the *r* corrected for research artifacts. We corrected the observed estimates as noted below. We were also careful to watch for other statistics (e.g., *d*, *t*, *F*, etc.) that could be converted to correlation coefficients. Nevertheless, all validity data from primary studies were reported as correlation coefficients.

We meta-analyzed our data with the Schmidt and Le (2004) program. We corrected correlation coefficients for analysis based on two different types of measures of job performance. For supervisory ratings of performance, we used the reliability of .60 (see Bobko et al., 1999 or Viswesvaran, Ones, & Schmidt, 1996, on this topic). For objective measures of performance, we used the reliability of .80 (e.g., Roth et al., 2003). We also looked for estimates of reliability for work sample tests as we coded. We found four test–retest reliability estimates for work samples (.76, .74, .71, and .61) and we used these as values when we wanted to correct for attenuation in the work sample exams.

Overlap with Previous Meta-Analyses

It is somewhat difficult to ascertain the exact overlap with previous meta-analyses. In two cases, there was not an explicit list of studies included in the work samples portion of the analysis (Hunter & Hunter, 1984; Schmitt et al., 1984). This is understandable as there were no conventions suggesting how authors communicate this information to readers in the earliest days of meta-analysis in applied psychology.

TABLE 1
Meta-Analysis of Work Sample Tests with Measures of Job Performance

Analysis	Mean *r*	Corrected mean *r*	*K*	*N*	Var.	80% CRI[1]	% SE
Overall	.26	.33	54	10,469	.005	.24–.42 (.24–.28)	55%
Overall w/*both* variables corrected	.26	.39	54	10,469	.007	.28–.50	55%

Note. Mean *r* is the observed *r* across studies, corrected mean *r* is the correlation corrected for unreliability in measures of job performance (and both work sample tests and measures of job performance are corrected for unreliability in line 2), *K* is the number of coefficients, *N* is the number of participants, Var. is the variance of the estimate of *rho* for the corrected coefficient without sampling error, CRI is an 80% credibility interval, and % SE is the percent of variance in validity coefficients explained by sampling error.

[1]95% confidence intervals are provided in parentheses beside the 80% credibility intervals.

We examined the overlap with Hunter (1983a) and we were able to retrieve data for 10 of the 14 coefficients from the articles and technical reports from this analysis. This is despite multiple attempts at interlibrary loans and other requests for documents. In terms of total samples, Hunter's work included 14 coefficients and 3,264 participants, and ours included 54 coefficients from approximately 10,000 participants.

Results

Coding Agreement

The authors coded a number of relatively continuous items and some dichotomous items from each study. In terms of continuous items, the coders' level of agreement is indicated by a correlation between their independent scores of .99 for the validity coefficient, .99 for sample size, and .77 for job complexity. Disagreements were resolved by discussion until consensus was reached. The two coders agreed 100% of the time for the objective versus subjective indicator, 100% for predictor versus criterion, and 100% for military versus nonmilitary. Coders agreed 98% of the time for applicant versus incumbent participants.

Meta-Analysis Results

Correlations with job performance. Our overall results in Table 1 indicate a work sample observed mean correlation of .26 ($K = 54$, $N = 10{,}469$). Correcting for criterion unreliability (assuming a value of

.6 for supervisory ratings and .8 for objective measures of performance) raises the mean correlation to .33. These results summarize data for validity studies in which the work sample, uncorrected for predictor unreliability, is correlated with job performance (i.e., one is conceptually interested in how work samples would predict job performance if job performance were measured with perfect reliability). We chose the value of .6 as an acceptable value for criterion reliability for subjective measures (Bobko et al., 1999; Viswesvaran et al., 1996) as well as to parallel the work of Hunter and Hunter (1984). It is also interesting that the 80% credibility interval for our overall analysis was from .24 to .42. Sampling error accounts for 55% of variance in validities.

One might also correct for unreliability in the work sample test as well as the measure of job performance. Conceptually, this would estimate how two such measures relate without the artifact of measurement unreliability. Correcting for attenuation in both measures led to a mean estimate of .39 (and an 80% credibility interval of .28 to .50). The point estimates of .33 and .39 are substantially lower than the value of .54 that is often cited as the benchmark for work sample validity for predicting job performance. The credibility intervals for both point estimates do not include the value of .54.

We also investigated several moderators. Our first moderator was the use of applicant versus incumbent participants. Analysis of such a moderating effect is not possible at this time as we found only one clear applicant study and N was only 24 (see Table 2). Perhaps one reason for this state of affairs is the time and expense of work sample test development and administration—there are simply fewer validity studies to be cumulated in this area as a whole relative to areas such as personality and cognitive ability. As such, there are also fewer applicant studies.

A second moderator, the use of objective versus subjective measures of job performance, does not appear to have a strong influence on work sample validity. We found that objective measures of job performance were associated with a mean observed validity of .27, though the K was only 8 and N was only 1,279. In a similar manner, subjective measures of job performance were associated with a mean correlation of .26 and data were more abundant ($K = 49$, $N = 9{,}339$). When corrected for attenuation in criteria (i.e., performance ratings and measures of output), the mean correlations rise to .34 and .30 for subjective and objective indicators of performance. Three studies provided both objective and subjective indicators of performance, and this allowed us to have a total of 57 coefficients for this analysis. Further, it is interesting to note that 39% and 58% of the variance in objective validities and subjective validities were accounted for by sampling error. Hence, there could be further moderators operating within these groups.

TABLE 2
Moderator Analysis of Work Sample Validity

Analysis	Mean *r*	Corrected mean *R*	*K*	*N*	Var.	80% CRI[1]	% SE
Applicants	.64	–	1	24	–	–	–
Incumbents	.26	.33	53	10,445	.005	.24–.42 (.24–.28)	57%
Objective	.27	.30	8	1,279	.011	.17–.43 (.19–.35)	39%
Subjective	.26	.34	49	9,339	.0057	.24–.43 (.24–.28)	58%
Predictor	.29	.37	24	1,741	.036	.13–.62 (.22–.36)	34%
Criteria	.25	.32	30	8,728	.000	.32–.32 (.23–.27)	100%
Military	.25	.32	24	7,295	.000	.32–.32 (.23–.27)	100%
Nonmilitary	.28	.35	30	3,174	.021	.16–.53 (.23–.33)	37%
Job complexity							
Low	.26	.31	4	942	.002	.26–.36 (.19–.33)	72%
Low medium	.28	.35	27	4,990	.010	.22–.48 (.24–.32)	41%
Medium	.25	.32	14	3,236	.000	.32–.32 (.22–.28)	100%
Medium high	.20	.25	6	1,030	.000	.25–.25 (.14–.26)	100%
1982 and before[2]	.31	.40	26	3,409	.012	.24–.56 (.26–.36)	40%
Post 1982	.25	.31	28	7,414	.003	.24–.39 (.22–.28)	59%

Note. Mean *r* is the observed *r* across studies, corrected mean *r* is the correlation corrected for unreliability in measures of job performance, *K* is the number of coefficients, *N* is the number of participants, Var. is the variance of the estimate of *rho* for the corrected coefficient without sampling error, CRI is an 80% credibility interval, and % SE is the percent of sampling error.

[1]95% confidence intervals are provided in parentheses below the 80% credibility intervals.

[2]We coded when studies were published by the date of publication. When there were multiple publication dates, we used the first publication date.

The distinction of using the work sample to serve as a predictor or criterion does not appear to be a particularly meaningful moderator. The predictor mean correlation was .29 ($K = 24$, $N = 1{,}741$) and the criterion mean correlation was .25 ($K = 30$, $N = 8{,}728$). Correcting for measurement error in the ratings of supervisor performance increased the estimates to .37 and .32, respectively.

There did not appear to be moderation of validity when comparing military to nonmilitary samples. The mean observed validity for military samples was .25 ($K = 24$, $N = 7{,}295$) versus .28 ($K = 30$, $N = 3{,}174$) for nonmilitary samples and corrections for unreliability did not increase any differences between military and nonmilitary samples.

Results for the complexity moderator were somewhat less straightforward as there was no easily interpretable linear trend in the data. The mean correlations for the complexity levels of low, low medium, and medium were rather similar in that all three mean observed correlations ranged from .25 to .28. The mean validity for medium high complexity was .20 though the sample size was only $K = 6$ and $N = 1{,}030$. We coded only 51 coefficients on this variable because studies that reported results across jobs were not coded if the jobs differed in complexity level. Further, we did not code studies for which they were not reasonably sure that enough information was provided upon which to base their judgments.

Table 2 also provides an analysis of validity coefficients from studies before and after 1982. Earlier studies were associated with an observed validity of .31 and a criterion corrected validity of .40. More recent studies were associated with an observed validity estimate of .25 corrected to .31.

We also report an analysis via correlation and regression, as suggested by a reviewer, of how our moderators related to each other. Table 3 (top panel) contains a column of how each moderator score was correlated with the validity of work sample exams. Immediately to the right of this, the intercorrelation matrix of the moderators is provided. Table 3 is based on 51 studies. Three studies were lost to listwise deletion.

Three correlations may be notable in the top panel of Table 3. The correlation of $-.51$ between complexity and subjectivity indicates that objective measures of performance tended to be associated with less complex jobs. The correlation of .47 between year and predictor indicates that work samples were more apt to be reported as criteria after 1982 and the correlation of $-.38$ between predictor and military suggests that military studies tended to use work samples more as criteria. Many of the correlations in Table 3 are either point-biserial correlations or *phi* coefficients.

Results for a multiple regression (predicting validity values from moderators) are also reported in Table 3. The confidence intervals for the regression coefficients all include zero. In terms of overall results, approximately 19% of the variance in observed validity is accounted for by the moderators. Hence, the moderators explain only a portion of the variance in the observed validity coefficients (though one should also recall that sampling error explains a nontrivial amount of this variation).

Correlations with other variables. As suggested by a reviewer, we also tried to cumulate correlations between three other predictors of job

TABLE 3
Correlations and Multiple Regression of Moderators for the Work Sample Test–Job Performance Relationship (N = 51 Studies)

	Validity	Appl.	Subjt.	Predictor	Military	Complex
Correlation Matrix						
Applicant	.32					
Subjective	.00	−.06				
Predictor	−.40	−.16	−.16			
Military	.18	.09	.03	−.38		
Complex	−.22	−.08	−.51	.13	−.14	
Year	−.33	−.15	−.03	.47	.00	.11

	Standardized weights	Unstandardized weights	95% confidence interval
Multiple Regression Weights			
Applicant	.227	.239	−.037–.515
Subjective	−.151	−.060	−.183–.062
Predictor	−.278	−.081	−.176–.013
Military	.017	−.001	−.088–.099
Complex	−.227	−.041	−.097–.014
Year	−.142	−.041	−.128–.046
Multiple Regression Results			
Multiple R of .53			
R^2 of .28			
Adjusted R^2 of .19			

Note. Validity refers to the observed/uncorrected validity, Appl. refers to applicant or incumbent status (0 = *incumbent*, 1 = *applicant*), Subjct. refers to subjective versus objective measures of job performance (0 = *subjective*, 1 = *objective*), Predictor refers to whether the authors conceptualized the work sample as a predictor or criterion (0 = *predictor*, 1 = *criterion*), Military refers to whether the sample was from a military or non-military population (0 = *military*, 1 = *non-military*), Complex refers to the complexity on a 1–5 scale with 5 being the most complex, Year refers to whether the study was published up to 1982 or post 1982 (0 = *up to1982*, 1 = *post 1982*).

performance and work sample exams (see Table 4). We found only one study that reported the relationship between a work sample test and a structured interview (see Pulakos & Schmidt, 1996). Nevertheless, we found 43 coefficients showing a relationship between a work sample and a measure of general cognitive ability. The mean observed correlation was .32 ($K = 43$, $N = 17{,}563$). The value rose to .38 when corrected for work sample unreliability and .40 when both types of tests were corrected for unreliability (we used a value of .90 for cognitive ability tests).

We conducted a number of moderator analyses that paralleled the form of previous moderator analyses and found that the observed correlations with cognitive ability were somewhat higher for work samples viewed as predictors versus criteria and higher for nonmilitary versus military

TABLE 4

Meta-Analysis of Correlations of Work Sample Tests With Cognitive Ability Tests and Situational Judgment Tests

Analysis	Mean *r*	Corrected mean *R*	*K*	*N*	Var.	80% CRI[1]	% SE
Measures of general cognitive ability							
Overall	.32	.38	43	17,563	.012	.24–.52 (.29–.35)	18%
Overall Construct	.32	.40	43	17,563	.013	.25–.55	20%
Predictor	.36	.43	7	1,082	.002	.37–.48 (.30–.42)	79%
Criterion	.32	.38	36	16,480	.013	.23–.52 (.29–.35)	18%
Military	.30	.36	27	12,524	.001	.24–.47 (.27–.33)	25%
Non-Military	.37	.44	16	5,039	.017	.29–.61 (.31–.43)	18%
Military w/ Range restr.	–	.48	14	6,100	.029	.26–.69	8%
Job complexity							
Low	.21	.25	2	842	.013	.11–.40 (.06–.36)	18%
Low Medium	.34	.40	22	8,682	.012	.27–.54 (.30–.38)	19%
Non-Military	.42	.50	7	2,503	.010	.37–.63 (.35–.49)	20%
Military	.31	.37	15	6,179	.007	.26–.47 (.27–.35)	27%
Medium	.28	.34	12	3,558	.014	.19–.50 (.22–.34)	22%
Non-Military	.42	.50	4	1,156	.000	.50–.50 (.38–.46)	100%
Military	.22	.26	8	2,402	.010	.19–.32 (.17–.27)	62%
Medium High	.34	.41	6	4,292	.003	.33–.48 (.29–.39)	30%
Situational judgment tests							
Overall	.13	–	3	1,571	.000	.13–.13 (.05–.21)	37%

Note. Mean *r* is the observed *r* across studies, corrected mean *r* is the correlation corrected for unreliability in work sample, *K* is the number of coefficients, *N* is the number of participants, Var. is the variance of the estimate of *rho* for the corrected coefficient without sampling error, CRI is an 80% credibility interval, and % SE is the percent of sampling error.

[1]95% confidence intervals are provided in parentheses below the 80% credibility intervals.

samples. We were also able to find 14 coefficients within the category of military studies that reported correlations that were corrected for range restriction. Nonetheless, some of the studies may have made corrections back to the general population (and not necessarily the applicant pop-

ulation). Interestingly, the correlation between work sample tests and measures of general cognitive ability was .48 when corrected for range restriction.

We also examined how job complexity moderated the relationship between cognitive ability and work samples. We found observed correlations in the .20s and .30s across all levels of complexity.

We further "broke down" our complexity results into military and nonmilitary samples. We did this because we are aware that the U.S. military uses a measure of cognitive ability in its selection process and this will likely result in a downward bias in estimates of *rho* due to direct range restriction. Therefore, this research artifact may heavily influence results in military samples. Interestingly, the observed correlations for low-medium complexity nonmilitary jobs and medium complexity nonmilitary jobs were .42 and both corrected to .50 when corrected for work sample unreliability (the correlation rose to .52 when corrected for unreliability in measures of general cognitive ability). Nevertheless, we caution readers that there were only seven and four studies in these categories. Other complexity-related results are also shown in Table 4.

In addition, recall that we found 14 coefficients above that used corrections for range restriction and the value was .48. All told, there may be a fairly strong relationship between work sample tests and measures of general cognitive ability. In many cases, the value (when not constrained by research artifacts) may be in excess of .50 for jobs of low-medium complexity and medium complexity jobs.

The observed relationship between work sample exams and situational judgment tests was .13 ($K = 3$, $N = 1{,}571$) in Table 4. The situational judgment tests (and work sample tests as well) may have been designed to measure a variety of constructs and results might vary depending upon the constructs that are targeted for measurement.

Incremental validity analyses. A reviewer strongly encouraged us to create a meta-analytic correlation matrix in order to regress job performance on both general cognitive ability and work sample tests (as per Schmidt & Hunter, 1998). Given the methodological limitations of available data (e.g., incapacity of the available data to account for differential range restriction), we tentatively offer our results. We used the value of .33 for the overall validity of work samples and .32 for the overall work sample—general cognitive ability correlations (see above). We also generated the value of .39 for the validity for general cognitive ability from the meta-analysis by Bobko et al. (1999). We used the observed value of .30 from Bobko et al., and corrected it for criterion unreliability using the value of .60 (again, we parallel the work of Hunter & Hunter, 1984).

TABLE 5
Meta-Analytic Matrix of Work Samples, General Cognitive Ability, and Job Performance

	Correlation Matrix		
	General cognitive ability	Work sample	Job performance
General cognitive ability			
Work sample	.32		
Job performance	.39[1]	.33	
Multiple regression results			
Multiple *R*	.45		
Change in *R*	.06		

[1]Adapted from Bobko, Roth, and Potosky (1999).

Both validities (.33 and .39) were corrected only for criterion unreliability but not for range restriction (because of a lack of such data in work sample tests). To maintain similarity, we also used the observed correlation between work samples and general cognitive ability.

The multiple *R* for our analysis in Table 5 was .45, and the incremental validity was .06 for adding a work sample test in addition to a test of general cognitive ability. We urge substantial caution in interpreting these results because (a) neither validity was corrected for range restriction and so both validities are biased downward and (b) there could be differential range restriction among the correlations in Table 5.

Discussion

Validity of Work Samples and Implications

The results of our analyses of work sample validity reinforce some beliefs, but, perhaps more important, our results may change other beliefs about the magnitude of the criterion related validity of work samples. In our overall analysis, we found that the observed (uncorrected) validity for work samples is .26 and the validity corrected for criterion unreliability in measures of job performance is .33. Thus, our results, along with those of Hunter and Hunter (1984; mean $r = .54$ corrected for criterion unreliability) and Schmitt et al. (1984; mean observed $r = .32$, $K = 7$, $N = 382$) all point towards evidence of validity for predicting job performance.

Although there is convergence on the fact that work samples can be valid predictors of job performance, our mean estimates are notably lower than some often cited previous estimates. Our overall validity of .33 is

conceptually comparable to Hunter and Hunter's estimate of .54 in that both are corrected only for measurement error in job performance. For instance, our overall estimate is .21 (or 39%) lower. An important implication of our review and systematic analysis is that the corresponding utility of work sample exams may be lower than previously thought.

We also examined several potential moderators in our up-dated meta-analysis. First, we found that there was only one coefficient from a study based on job applicants so no meaningful moderator analyses could be performed. We also found that results for objective and subjective criteria did not appear to have any notable moderating influence—the observed correlations were within .01 of each other, and the corrected correlations were .04 apart. We believe ours is the first meta-analysis to conduct such analyses for work samples, and thus, we add this understanding to the literature.

There also did not appear to be a moderating effect in the work sample–job performance correlations due to considering the work sample a predictor or a criterion. Studies in which the work samples were conceptualized as a predictor were associated with .05 larger corrected correlations with measures of job performance than studies in which the work samples were conceptualized as a criterion. Nevertheless, the credibility intervals also overlapped. We also found little effect due to military versus nonmilitary status on validity. Although Hunter (1983a) found a somewhat marked moderating effect in this case, his sample sizes were rather small (and he noted this limitation in his own work). Our larger sample size allows us to update findings in this area. Finally, we did not find a clear linear trend that could characterize the effect of job complexity on validities.

We also examined an effect for the year a study was published. There is evidence to suggest that studies published after 1982 are associated with a somewhat lower mean validity (.25) than studies published up to 1982 (.31), and the corrected values increased to .31 and .40, respectively. Therefore, there is some evidence that older studies were associated with higher observed validities, and this may partially explain differences between our results and those of Hunter and Hunter (1984).

Our analyses also indicate that the work of Asher and Sciarrino (1974), which has been influential in the understanding of work sample–performance relationships, is filled with methodological problems. Hence, we would urge caution in using any later meta-analytic re-analysis of such work (recall some of Hunter's own cautions about this database as well, e.g., see Hunter & Hunter, 1984, p. 84).

To illustrate our reasons for caution, and to try to shed some light on why we found differences with previous analyses, we examined six coefficients that were associated with criterion contamination. Our bare-bones

meta-analysis of these six coefficients resulted in a mean observed correlation of .69 ($K = 6$, $N = 128$) between work sample exams and criteria of job performance and training success. Such a large correlation is in the direction we would expect, although this single methodological problem alone is not likely to have seriously influenced previous results (because of small N). We also tried to conduct similar analyses for studies that suffered from range enhancement but could not conduct analyses due to reporting limitations in the results of these studies.

Our analysis also considered relationships between work samples and two other types of predictors of job performance. First, there appears to be at least a moderate correlation between work sample tests and measures of general cognitive ability. Our overall estimate of .32 is probably downwardly biased (i.e., conservative) as it is likely influenced by range restriction. A number of military studies that corrected for range restriction suggested the relationship could be .48. So, unbiased estimates of *rho* could be higher when all research artifacts are taken into account. These meta-analytic estimates address the need suggested by other researchers (Hunter & Hunter, 1984; Bobko et al., 1999) for studies of predictor intercorrelations. Second, work sample scores did not appear to correlate highly with situational judgment tests. The observed correlation was .13 but was based on only three coefficients.

We also conducted an analysis on a meta-analytic matrix in which we regressed job performance on work samples and a measure of general mental ability. Incremental validity for the work sample test was .06. We did not place a great deal of weight on our results, given that we could not correct work sample tests and general cognitive ability for range restriction.

Limitations

All studies have limitations and we mention four. First, we reiterate the lack of any information on the amount of range restriction in any of our validity studies. Although we looked for such information, we were simply unable to find it. The lack of a correction for range restriction makes comparisons between work samples and other predictors difficult. Second, we found relatively few studies that reported reliabilities. We found only four test–retest reliabilities and we used these in our corrections for work sample test validity when appropriate.

Third, we also point out our moderate sample size relative to other existing meta-analyses of validities of other predictors of job performance (e.g., Hunter & Hunter's, 1984, analysis of 515 coefficients for the validity of cognitive ability tests). One reason for the moderate sample size may

be the expense of test development and administration. Finally, we were unable to find data to code and report validity results by dimensions within a work sample exam. For example, a work sample test may have been made up of multiple exercises/components such as a scheduling exercise, an exercise in which participants interacted with customers or coworkers, and so forth. It had been our intent to look at these dimensions and examine them as well as overall scores.

Future Research

There are several areas within the work sample literature that could benefit from future research. Conducting predictive validity studies is important for both informed decisionmaking about selection systems and legal considerations. In terms of informed decisionmaking, applied psychologists and human resource management professionals often compare selection devices on attributes such as validity, adverse impact, and feasibility as they determine which devices to use in a selection system. Validity estimates corrected for range restriction are available for some selection devices (e.g., tests of cognitive ability, interviews). Having corrected validity estimates for work sample tests would allow decisionmakers to compare work sample tests to other possible tests without the confounding influence of differential range restriction as a research artifact. In terms of legal considerations, the *Uniform Guidelines* notes that if a test shows adverse impact, organizations should generally use other tests with lower levels of adverse impact as long as the alternative tests have substantially equal validity to the original. As mentioned above, range restriction corrected validity estimates would be helpful in addressing this issue.

The issue of range restriction also has important implications for the study of standardized ethnic group differences (d) on work sample exams. Although we explicitly looked for ds computed on job applicants in this project, we were disappointed and unsuccessful in finding such statistics. This is because virtually all of the estimates of standardized ethnic group differences for Blacks versus Whites (and Hispanics vs. Whites) in the work sample literature were based on studies of incumbents (e.g., see also Roth et al., 2003, for a meta-analytic review of this data). This state of affairs likely does not inform decisionmakers of the expected level of adverse impact that organizations might expect at the beginning of the selection process if they use such exams. In fact, one could argue that using results based on samples of job incumbents should result in range restriction on estimates of d and systematically bias these statistics downward. In this case, the d associated with job applicants would likely be higher and organizational decisionmakers using work sample tests may

be surprised by higher than expected actual levels of adverse impact. This is particularly salient in light of our finding that the validity of work sample tests may be lower than previously thought.

We also believe that work sample tests could benefit from a continued focus on the constructs they assess. Sackett, Schmitt, Ellingson, and Kabin (2001) narratively review a great deal of selection literature and reaffirm that various predictors of job performance, such as work sample exams, can measure a variety of constructs. One approach to examining constructs in work sample tests is to look at exercises or dimensions within work sample exams. Again, we looked for this information during our search, but it was seldom reported in the existing literature. We suggest the approach of looking at individual exercise/component scores for two reasons. First, work sample exams often have more than one exercise/component or dimension. For example, there might be a technical exercise in some area (e.g., finance) and an exercise on counseling a subordinate. The overall score for the work sample might be computed by summing the two exercise/component scores in some way. Second, focusing on the individual exercise scores might also allow researchers to "disaggregate" the KSAs from the exam level back to the exercise level in order to get a somewhat clearer picture of what is being measured.

We also call for future research on the fidelity of predictors of job performance. Work sample tests might generally fall along the high end of the fidelity range but other predictors such as situational judgment tests and perhaps some interview questions might fall along the middle to lower end of the fidelity range. Analyses might then compare the validity of various levels of fidelity. We also suggest that future researchers consider the role of general cognitive ability in such analyses such that they code for the saturation of this construct.

In conclusion, we meta-analyzed work sample tests in order to update some classic literature (Asher & Sciarrino, 1974; Hunter & Hunter, 1984; Schmitt et al., 1984), as well as thoroughly screen previous work to make sure we were analyzing only work sample exams from studies without major methodological problems (e.g., inclusion of nonwork sample exams). We believe there is evidence for the validity of work sample exams and many interesting research projects in this area. On the other hand, our extensive meta-analysis found that the point estimate of work sample validity is substantially lower than generally stated by researchers and practitioners in the field. Work sample tests have appeal for a variety of reasons described above (e.g., favorable applicant reactions, potential to link to content of a job, etc.), but some organizations may be overestimating the validity and utility of such selection devices. We look forward to the increased understanding that future research on this topic will bring.

REFERENCES

*Denotes studies that contained data used in the meta-analyses.

Abt LEA. (1949). A test battery for selecting technical magazine editors. PERSONNEL PSYCHOLOGY, *2*, 75–19.

Adams WM. (1948). Prediction of scholastic success in colleges of law. *Educational & Psychological Measurement, 3*, 291–305.

*Anonymous. (1954). Validity information exchange No. 7-094. PERSONNEL PSYCHOLOGY, *7*, 572.

Ash P. (1960). Validity information exchange No. 13-07. PERSONNEL PSYCHOLOGY, *13*, 456.

Asher JJ, Sciarrino JA. (1974). Realistic work sample tests: A review. PERSONNEL PSYCHOLOGY, *27*, 519–533.

*Avis JM. (2001). *An examination of the prediction of overall, task, and contextual performance using three selection measures for a service-type occupation*. University of Southern Mississippi, Unpublished Doctoral Dissertation.

Baier DE, Dugan RD. (1956). Tests and performance in a sales organization. PERSONNEL PSYCHOLOGY, *9*, 17–26.

Bass BM. (1954). The leaderless group discussion as a leadership evaluation instrument. PERSONNEL PSYCHOLOGY, *7*, 470–477.

Bass BM, Coates CH. (1954). Validity information exchange No. 7-082. PERSONNEL PSYCHOLOGY, *7*, 553–554.

Bender WRG, Loveless HE. (1958). Validation studies involving successive classes of trainee stenographer. PERSONNEL PSYCHOLOGY, *11*, 491–508.

*Bennett GK, Fear RA. (1943). Mechanical comprehension and dexterity. *Personnel Journal, 22*, 12–17.

Blum ML. (1943). Selection of sewing machines operators. *Journal of Applied Psychology, 27*, 35–40.

Bobko P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). London: Sage.

Bobko P, Roth PL, Potosky D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors and job performance. PERSONNEL PSYCHOLOGY, *52*, 561–589.

Bommer WH, Johnson J, Rich GA. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. PERSONNEL PSYCHOLOGY, *48*, 587–605.

Callahan DK. (1993). Work hardening for a client with low back pain. *American Journal of Occupational Therapy, 47*, 645–649.

*Campion JE. (1972). Work sampling for personnel selection. *Journal of Applied Psychology, 56*, 40–44.

Callinan M, Robertson IT. (2000). Work sample testing. *International Journal of Selection & Assessment, 8*, 248–260.

Carpenter CR, Greenhill LP, Hittinger WF, McCoy EP, McIntyre CJ, Murnin JA, et al. (1954). The development of sound motion picture proficiency test. PERSONNEL PSYCHOLOGY, *7*, 509–523.

Cascio W. (2003). *Managing human resources: Productivity, quality of work life, and profits* (6th Edition). Boston: McGraw-Hill.

*Clevenger J, Pereira GM, Wiechman E, Schmitt N, Harvey VS. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417.

Davis FB (Ed.). (1947). *Army air forces aviation psychology program research reports*. The AAF Qualifying Examination Report No. 6, Chapter 6. Washington, DC: Government Printing Office.

Drewes DW. (1961). Development and validation of synthetic dexterity tests based on elemental motion analysis. *Journal of Applied Psychology, 45*, 179–185.

*Dubois CLZ, Sackett PR, 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.

DuBois PH, Watson RI. (1950). The selection of patrolmen. *Journal of Applied Psychology, 34*, 90–95.

*Ekberg DL. (1947). A study in tool usage. *Educational and Psychological Measurement, 7*, 421–427.

*Engel JD. (1970). *Development of a work sample criterion for general vehicle mechanic*. Fort Knox, KY: HumRRO for Research and Development Department of the Army (Report 70-11).

*Fattu NA, Pfeiffer EL, Demaree RG, Wilder CE. (undated). *An evaluation of selected machinist tests for possible use as Air Force machinist proficiency measures*. Chanute AFB, IL: Training Aids Laboratory (project 507-012-0002).

*Field HS, Bayley GA, Bayley SM. (1977). Employment test validation for minority and nonminority production workers. PERSONNEL PSYCHOLOGY, *30*, 37–46.

Forehand GA, Guetzkow H. (1961). The administration judgment test as related to description of executive judgement behaviors. *Journal of Applied Psychology, 45*, 257–261.

Franze IJ, Ferrari JR. (2002). Career search efficacy among an at-risk sample: Examining changes among welfare recipients. *Journal of Prevention and Intervention in the Community, 23*, 119–128.

*Gael S, Grant DL. (1972). Employment test validation for minority and non-minority telephone company service representatives. *Journal of Applied Psychology, 56*, 135–139.

*Gael S, Grant DL, Ritchie RJ. (1975). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology, 60*, 420–426.

*Gael S, Grant DL, Ritchie RJ. (1975). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60*, 411–419.

Gatewood RD, Field HS. (2001). *Human Resource selection (5th ed.)*. New York: Harcourt.

*Glanz E. (1949). A trade test for power sewing machine operators. *Journal of Applied Psychology, 33*, 436–441.

*Giese WJ. (1949). A tested method for the selection of office personnel. PERSONNEL PSYCHOLOGY, *2*, 525–545.

Glaser R, Schwarz PA, Flanagan JC. (1958). The contribution of interview and situational performance procedures to the selection of supervisory personnel. *Journal of Applied Psychology, 42*, 69–73.

Gleason WJ. (1957). Predicting Army leadership ability by modified leaderless group discussion. *Journal of Applied Psychology, 41*, 231–235.

*Grant DL, Bray DW. (1970). Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology, 54*, 7–14.

Grigg AE. (1948). A farm knowledge test. *Journal of Applied Psychology, 32*, 452–455.

Guion RM. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Handyside JD, Duncan DC. (1954). Four years later: A follow-up of an experiment in selecting supervisors. *Occupational Psychology, 28*, 9–23.

Harville DL. (1996). Ability test equity in predicting job performance work samples. *Educational and Psychological Measurement, 56*, 344–348.

*Hattrup K, Schmitt N. (1990). Prediction of trades apprentices' performance on job sample criteria. PERSONNEL PSYCHOLOGY, *43*, 453–466.

Hausknecht JP, Day DV, Thomas SC. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. PERSONNEL PSYCHOLOGY, *57*, 639–683.

Hedge JW, Teachout MS. (1992). An interview approach to work sample criterion measurement. *Journal of Applied Psychology, 77*, 453–461.

*Hedge JW, Teachout MS, Laue FJ. (1990). *Interview testing as a work sample measure of job proficiency*. Brooks, AFB: Training Systems Division, Training Assessment Branch (AFHRL TP 90-61).

Heneman RL. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. PERSONNEL PSYCHOLOGY, *39*, 811–826.

Heneman HG, Judge T. (2003). *Staffing organizations* (4th ed.). Irwin/McGraw-Hill.

Highhouse S, Gallo A. (1997). Order effects in personnel decision making. *Human Performance, 10*, 31–46.

*Hollenbeck GP, McNamara WJ. (1965). CUCPAT and programming aptitude. PERSONNEL PSYCHOLOGY, *18*, 101–106.

Hough L. (1998). Personality at work: Issues and evidence. In Hakel M (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131–159). Mahwah, NJ: Erlbaum.

Hough LM. (1984). Development and evaluation of the accomplishment record method of selecting and promoting professionals. *Journal of Applied Psychology, 69*, 135–146.

Hunter JE. (1983a). A causal analysis of cognitive ability, job knowledge, and job performance, and supervisory ratings. In Landy R, Zedeck S, Cleveland J (Eds.), *Performance measurement and theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.

Hunter JE. (1983b). *Test validation for 12,000 jobs: An application of job classification and validity generalization to the General Aptitude Test Battery (GATB)*. USES Test Research Report No. 45. Washington DC: U.S. Department of Labor.

Hunter JE. (1986). Cognitive ability, cognitive aptitude, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340–362.

Hunter JE, Hunter RF. (1983). *The validity and utility of alternative predictors of job performance*. Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development (OPRD-83-4).

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter JE, Schmidt FL. (1990). *Methods of meta-analysis: Correcting for error and bias in research findings*. Newbury Park, CA: Sage.

Hunter JE, Schmidt FL. (2004). *Methods of meta-analysis: Correcting for error and bias in research findings* (2nd Edition). Newbury Park, CA: Sage.

*Inskeep GC. (1971). The use of psychomotor tests to select sewing machine operators some negative findings. PERSONNEL PSYCHOLOGY, *24*, 707–714.

*Jackson DN, Harris WG, Ashton MC, McCarthy JM, Tremblay PF. (2000). How useful are work samples in validation studies? *International Journal of Selection and Assessment,* 8, 29–33.

Kazmier LJ. (1962). Criterion simulation and the prediction of achievement. *Psychological Reports, 10*, 64.

Knauft EB. (1949). A selection battery for bake shop managers. *Journal of Applied Psychology, 33*, 304–315.

Kriedt PH. (1952). Validation of a correspondence aptitude test. *Journal of Applied Psychology, 36*, 5–7.

Latham GP, Saari LM, Campion ME, Pursell ED. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

Lyons TJ, Bayless JA, Park RK. (2001). Relationship of cognitive, biographical, and personality measures with the training and job performance of detention enforcement officers in a federal agency. *Applied HRM Research, 6*, 67–70.

Mandell MM. (1947). The selection of foreman. *Educational and Psychological Measurement, 7*, 385–397.

Mandell MM. (1950). The administrative judgment test. *Journal of Applied Psychology, 34*, 145–147.

*Manning CA. (2000). *Measuring air traffic controller performance in a high fidelity simulation*. Washington, DC: U.S. Department of Transportation (Report DOT/FAA/AM-00/2).

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

McEvoy GM, Cascio WF. (1989). Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology, 74*, 11–17.

McNamara WJ, Hughes JL. (1961). A review of research on the selection of computer programmers. PERSONNEL PSYCHOLOGY, *14*, 39–51.

*Melvin KB, Haigler MH, Sims LJ, McDowell DJ. (1994). Validation of the Melvin-Simms word processing operator test. *Journal of Business & Psychology, 9*, 199–221.

*Meyer HH. (1970). The validity of the in-basket test as a measure of managerial performance. PERSONNEL PSYCHOLOGY, *23*, 297–307.

Motowidlo S, Dunnette MD, Carter GW. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640–647.

Mount MK, Muchinsky PM, Hanser LM. (1977). The predictive validity of a work sample: A laboratory study. PERSONNEL PSYCHOLOGY, *30*, 637–645.

Murphy KR, Cleveland JN. (1991). *Performance appraisal*. Needham Heights, MA: Allyn & Bacon.

Nathan BR, Alexander RA. (1988). A comparison of criteria for test validation: A meta-analytic investigation. PERSONNEL PSYCHOLOGY, *41*, 517–535.

Olea MM, Ree MJ. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79*, 845–851.

*O'Leary BS, Trattner MH. (1977). *Research base for the written portion of the professional and administrative career exam (PACE): Prediction of performance for internal revenue officers*. Washington, DC: Personnel Research and Development Center.

Ones DS, Viswesvaran C, Schmit FL. (1993). Comprehensive meta-analysis of integrity test validation: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703.

*Palmer CI, Boyles WR, Veres JG, Hill JB. (1992). Validation of a clerical test using work samples. *Journal of Business & Psychology, 7*, 239–257.

Perone M, DeWaard RJ, Baron A. (1979). Satisfaction with real and simulated jobs in relation to personality variables and drug use. *Journal of Applied Psychology, 64*, 660–668.

Ployhart RE, Schneider B, Schmitt N. (in press). *Staffing organizations: Contemporary practice and research*. Hillsdale or Mahwah, NJ: Erlbaum.

Poruben A. (1950). A test battery for actuarial clerks. *Journal of Applied Psychology, 34*, 159–162.

Pritchard RD, Hollenback J, DeLeo PJ. (1980). The effects of continuous and partial schedules of reinforcement on effort, performance, and satisfaction. *Organizational Behavior and Human Decision Processes, 25*, 336–353.

*Pulakos ED, Schmitt N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*, 241–258.

*Ree ML, Carretta TR, Teachout MS. (1995). Role of ability and prior job knowledge in complex job training. *Journal of Applied Psychology, 80*, 721–730.

Reilly RR, Warech MA. (1993). The validity and fairness of alternatives to cognitive ability tests. In Wing L, Gifford B (Eds.), *Policy issues in employment testing*. Boston: Kluwer.

Robertson IT, Mindel RM. (1980). A study of trainability testing. *Journal of Occupational Psychology, 53*, 131–138.

Roth PL, Huffcutt AI, Bobko P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*, 725–740.

Russell CJ, Dean MA. (1994, month). *The effect of history on meta-analytic results: An example from personnel selection research*. Presented at the Annual Meeting of the Academy of Management, Dallas, TX.

Sackett PR, Schmitt N, Ellingson JE, Kabin ME. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post affirmative action world. *American Psychologist, 56*, 302–318.

Salgado JF, Viswesvaran C, Ones DS. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In Anderson N, Ones D, Sinangil H, Viswesvaran C (Eds.), *Handbook of industrial, work, and organizational psychology* (pp. 165–199). London: Sage.

Schmidt FL, Hunter JE. (1998). The validity of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt FL, Le H. (2004). *Software for the Hunter-Schmidt meta-analysis methods*. University of Iowa, Department of Management & Organization. Iowa City, 42242.

Schmidt FL, Hunter JE, Outerbridge AN, Goff S. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology, 73*, 46–57.

Schmitt N, Clause CS, Pulakos ED. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In Cooper CL, Robertson (Eds.), *International review of industrial and organizational psychology* (vol. 11, pp. 115–139).

Schmitt N, Gooding RZ, Noe RA, Kirsch M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. PERSONNEL PSYCHOLOGY, *37*, 407–422.

*Schmitt N, Mills AE. (2001). Traditional test and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86*, 451–458.

Schneider B, Schmitt N. (1986). *Staffing organizations*. Glenview, IL: Scott Foresman & Co.

*Schoon CG, Kaley M, Stern I. (1979). *Correlation of performance on clinical laboratory proficiency examination with performance in clinical laboratory practice*. Myattsville, MD: Department of Health, Education, & Welfare (Contract HRA-231-777-0018).

*Scott TM. (1985). *Relative validity and utility of background data, work samples, and cognitive tests as predictors of data entry clerk performance*. Dissertation from Georgia State University (Atlanta, GA).

*Seashore HG, Bennett GK. (1948). A test of stenography: Some preliminary results. PERSONNEL PSYCHOLOGY, *1*, 197–209.

Terpstra DE, Kethley RB, Foley RT. (2000). The nature of litigation surrounding five screening devices. *Public Personnel Management, 29*, 43–54.

*Thumin FJ. (1993). Predictor validity as related to criterion relevance, restriction of range, and ethnicity. *Journal of Psychology, 127*, 553–563.

*Uhlmann FW. (1962). A selection test for production machine operators. PERSONNEL PSYCHOLOGY, *15*, 287–293.

Vance R, Coovert M, MacCallum R, Hedge J. (1989). Construct models of task performance. *Journal of Applied Psychology, 74*, 447–455.

*Vineberg R, Taylor EN. (1972). *Performance in four army jobs by men at different aptitude (AFQT) levels: The relationship of AFQT and job experience to job performance*. Alexandria, VA: Human Resources Research Organization.

Viswesvaran C, Ones DS, Schmidt FL. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Weislogel RL. (1954). Development of situational tests for military personnel. PERSONNEL PSYCHOLOGY, *7*, 492–497.

West L, Bolanovich DJ. (1963). Evaluation of typewriting proficency: Preliminary test development. *Journal of Applied Psychology, 47,* 403–407.

*Wigdor AK, Green BF. (1991). *Performance assessment for the workplace* (vol. 1). Washington DC: National Academy Press.

Wernimont PF, Campbell JP. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376.