The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years...

Working Paper · October 2016

DOI: 10.13140/RG.2.2.18843.26400

1 author:

Frank L. Schmidt
University of Iowa
222 PUBLICATIONS **26,512** CITATIONS

SEE PROFILE

Running Head: Validity and Utility of Selection Methods

Working Paper

The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years of Research Findings

Frank L. Schmidt

University of Iowa

In-Sue Oh

Temple University

Jonathan A. Shaffer

West Texas A&M University

*** This working paper was prepared as an update to Schmidt and Hunter (1998). We have made it available on Research Gate and SSRN because of repeated requests for such updates from practitioners and researchers from all over the world. We intend to publish a version of this paper later.**

Authors' Note. Frank L. Schmidt, Department of Management and Organizations, University of Iowa, Iowa City, IA 52242, and Gallup Organization; In-Sue Oh, Department of Human Resource Management, Fox School of Business, Temple University, Philadelphia, PA 19122. Email: insue.oh@temple.edu; Jonathan Shaffer, Department of Management Marketing, and General Business, College of Business, West Texas A&M University. Email: jshaffer@wtamu.edu. Enquiries concerning this paper should be sent to Frank L. Schmidt, Department of Management and Organizations, University of Iowa, Iowa City 52242. Email: frank-schmidt@uiowa.edu

Abstract

This paper summarizes the practical and theoretical implications of 100 years of research in personnel selection. On the basis of meta-analytic findings, this paper presents the validity of 31 procedures for predicting job performance and the validity of paired combinations of general mental ability (GMA) and the 29 other selection procedures. Similar analyses are presented for 16 predictors of performance in job training programs. Overall, the two combinations with the highest multivariate validity and utility for predicting job performance were GMA plus an integrity test (mean validity of .78) and GMA plus a structured interview (mean validity of .76). Similar results were obtained for these two combinations in the prediction of performance in job training programs. A further advantage of these two combinations is that they can be used for both entry level hiring and selection of experienced job applicants. The practical utility implications of these summary findings are substantial. The implications of these research findings for the development of theories of job performance are discussed.

Keywords*: personnel selection, meta-analysis, validity generalization, selection utility, job performance.*

This paper is an update of Schmidt and Hunter (1998), which summarized 85 years of research findings on the validity of job selection methods up to 1998. That article that has been cited over 3,400 times, suggesting that an update of findings would be of interest to researchers, practitioners, and employers. The ensuing period of nearly 20 years has indeed seen methodological developments and additional research findings that refine and improve the accuracy of the validity estimates presented in the 1998 article. During this time, a new and more accurate procedure for correcting for the downward bias caused by range restriction has become available (Hunter, Schmidt, & Le, 2006). This more accurate procedure has revealed that the older, less accurate procedure had substantially underestimated the validity of general mental ability (GMA) and specific cognitive aptitudes (e.g., verbal ability, quantitative ability, etc.; Schmidt, Oh, & Le, 2006). Also, the increased availability of primary validity studies has allowed new and expanded meta-analyses of some selection methods, refining and changing some of the validity estimates for the prediction of job performance. For some personnel measures, these new data have produced important changes in estimated validity and incremental validity over GMA. For example, an expanded meta-analysis shows that job sample or work sample tests are somewhat less valid than had been indicated by the older data. Also, meta-analytic results are now available for some newer predictors not included in the 1998 article. These include Situational Judgment Tests (SJTs), college and graduate school grade point average (GPA), phone-based structured employment interviews, measures of "emotional intelligence", person-job fit measures, person-organization fit measures, and self-report measures of the Big Five personality traits.

We present the 31 personal selection procedures used to predict job performance in the order of their incremental validity (if any) produced over that of GMA. We also present the

mean zero order operational validity of each procedure as revealed by meta-analyses. We present this information for 16 procedures used to predict performance in job training programs in the same manner. Results show that many procedures that are valid predictors of job performance nevertheless have little or no incremental validity over that of GMA. The rank order for zero order validity is different from the rank order for incremental validity. Also, the incremental validity of most procedures is smaller than reported in Schmidt and Hunter (1998). This reduction in apparent incremental validity results from the increase in the estimated validity of GMA resulting from use of the more accurate correction for range restriction (Hunter et al., 2006; Hunter & Schmidt, 2004; Schmidt & Hunter, 2015). At the time of the earlier 1998 article, it was apparent that GMA plays a central role in the determination of both job and training performance. However, the more accurate updated findings indicate that the dominance of GMA is greater than previously believed.

The nature of the improvement in the correction for range restriction requires comment. Up until about 2006, all corrections for range restriction were based on the assumption of direct range restriction; that is, on the assumption of direct truncation on the predictor scores (i.e., all above the cut score were hired and all below were rejected). The correction used was Thorndike's Case II; Thorndike, 1949). It has long been known that this assumption was false and that virtually all range restriction in hiring was indirect in nature. That is, it was known that applicants were almost always hired based on a combination or composite of different factors and this composite was correlated with the selection method being studied, creating indirect range restriction on it. Hunter et al. (2006) presented a method for correcting for indirect range restriction that was widely usable because it did not require the typically unavailable information required to implement the older correction for indirect range restriction (Thorndike's Case III;

Thorndike, 1949). Simulation studies showed this procedure was appreciably more accurate than Thorndike's Case II (Le & Schmidt, 2006), and revealed that Thorndike's Case II substantially underestimated validity in the presence of indirect range restriction.[1]

From the point of view of practical value, the most important property of a personnel assessment method is predictive validity: the ability to predict future job performance, job-related learning (such as amount learned in training and development programs), and other criteria. The predictive validity coefficient is directly proportional to the practical economic value (utility) of the assessment method (Brogden, 1949; Schmidt, Hunter, McKenzie, & Muldrow, 1979). Use of hiring methods with increased predictive validity leads to substantial increases in employee performance as measured in percentage increases in output, increased monetary value of output, and increased learning of job-related skills (Hunter, Schmidt, & Judiesch, 1990).

Today, the validity of different personnel measures can be calibrated via the application of meta-analysis to 100 years of research studies. The most well-known conclusion from this research is that for hiring employees without previous experience in the job the most valid predictor of future performance and learning is general mental ability (GMA, i.e., intelligence or general cognitive ability; Brown, Le, & Schmidt, 2006; Hunter & Hunter, 1984; Hunter et al., 2006; Ree & Earles, 1992; Schmidt, Shaffer, & Oh, 2008). GMA can be measured using commercially available tests. However, many other measures can also contribute to the overall validity of the selection process. These include, for example, employment interviews and measures of conscientiousness and personal integrity.

This paper examines and summarizes what nearly 100 years of research in personnel psychology has revealed about the validity of measures of 31 different selection methods that can

be used in making decisions about hiring, training, and developmental assignments. In addition, this paper examines how well certain combinations of these methods work. These 31 procedures do not all work equally well; the research evidence indicates that some work very well and some work very poorly. Measures of GMA work very well and graphology, for example, does not work at all. The cumulative findings show that the research knowledge now available makes it possible for employers today to substantially increase the productivity, output, and learning ability of their workforces by using procedures that work well and by avoiding those that do not. Making this information available to employers and practitioners is important in light of research showing that the gap between research findings and real world practices is greater in the selection and staffing area than in any other area of human resource management (Rynes, Colbert & Brown, 2002; Rynes, Giluk, & Brown, 2007). Finally, we look at the implications of these research findings for the development of theories of job performance.

Determinants of Practical Value (Utility) of Selection Methods

The validity of a hiring method is a direct determinant of its practical value, but it is not the only determinant. Another direct determinant is the variability of job performance. At one extreme, if variability were zero, then all applicants would have exactly the same level of later job performance if hired. In this case, the practical value or utility of all selection procedures would be zero. In such a hypothetical case, it does not matter who is hired, because all workers are the same. At the other extreme, if performance variability is very large, it then becomes important to hire the best performing applicants and the practical utility of valid selection methods is very large. As it happens, this "extreme" case appears to be the reality for most jobs. Research has shown that the variability of performance and output among (incumbent) workers is very large and that it would be even larger if all job applicants were hired or if job applicants

were selected randomly from among those that apply (cf. Hunter et al., 1990; Schmidt & Hunter, 1983; Schmidt et al., 1979). This latter variability is called the applicant pool variability, and in hiring this is the variability that operates to determine practical value. This is because one is selecting new employees from the applicant pool, not from among those already on the job in question.

The variability of employee job performance can be measured in a number of ways, but two scales have typically been used: dollar value of output and output as a percentage of mean output. The standard deviation across individuals of the dollar value of output (called SD_y) has been found to be at minimum 40% of the mean salary of the job (Schmidt & Hunter, 1983; Schmidt et al., 1979; Schmidt, Mack, & Hunter, 1984). The 40% figure is a lower bound value; actual values are typically considerably higher. Thus, if the average salary for a job is $40,000, then SD_y is at least $16,000. If performance has a normal distribution, then workers at the 84th percentile produce output worth $16,000 more per year than average workers (i.e., those at the 50th percentile). And the difference between workers at the 16th percentile ("below average" workers) and those at the 84th percentile ("superior" workers) is twice that: $32,000 per year. Such differences are large enough to be important to the economic health of an organization.

Employee output can also be measured as a percentage of mean output; that is, each employee's output is divided by the output of workers at the 50th percentile and then multiplied by 100. Research shows that the standard deviation of output as a percentage of average output (called SD_p) varies by job level. For unskilled and semi-skilled jobs, the average SD_p figure is 19%. For skilled work, it is 32%, and for managerial and professional jobs, it is 48% (Hunter et al., 1990). These figures are averages based on all available studies that measured or counted the amount of output for different employees. If a superior worker is defined as one whose

performance (output) is at the 84th percentile (that is, 1 *SD* above the mean), then a superior worker in a lower level job produces 19% more output than an average worker, a superior skilled worker produces 32% more output than the average skilled worker, and a superior manager or professional produces output 48% above the average for those jobs. These differences are substantial and they indicate that the payoff from using valid hiring methods to predict later job performance is quite large.

Another determinant of the practical value of selection methods is the selection ratio—the proportion of applicants who are hired. At one extreme, if an organization must hire all who apply for the job, no hiring procedure has any practical value. At the other extreme, if the organization has the luxury of hiring only the top scoring 1%, the practical value of gains from selection per person hired will be extremely large. But few organizations can afford to reject 99% of all job applicants. Actual selection ratios are typically in the .30 to .70 range, a range that still produces substantial practical utility.

The formula for computing practical gains per person hired per year on the job is a three-way product (Brogden, 1949; Schmidt et al., 1979):

$$\Delta\bar{U}/hire/year = \Delta r_{xy} SD_y \bar{Z}_x$$

(when performance is measured in dollar value) (1)

$$\Delta\bar{U}/hire/year = \Delta r_{xy} SD_p \bar{Z}_x$$

(when performance is measured in percentage of average output) (2)

In these equations, Δr_{xy} is the difference between the validity of the new (more valid) selection procedure and the old selection procedure. (Both the new and the old selection procedures can be composites of scores on several selection methods.) If the old selection method has no validity (that is, selection is random), then Δr_{xy} is the same as the validity of the new procedure; that is,

$\Delta r_{xy} = r_{xy}$. Hence, relative to random selection, practical value (utility) is directly proportional to validity. If the old procedure has some validity, then the utility gain is directly proportional to Δr_{xy}. The term $\bar{Z}_x$ is the average score on the employment procedure of those hired (in z -score form), as compared to the general applicant pool. The smaller the selection ratio, the higher this value will be. The first equation expresses selection utility in dollars. For example, a typical final figure for a medium complexity job might be $18,000, meaning that increasing the validity of the hiring methods leads to an average increase in output per hire of $18,000 per year. To get the full value, one must of course multiply by the number of workers hired. If 100 are hired, then the increase would be (100)($18,000) = $1,800,000. Finally, one must consider the number of years these workers remain on the job, because the $18,000 per worker is realized each year that worker remains on the job. Of all these factors that affect the practical value, only validity is a characteristic of the personnel measure itself.

The second equation expresses the practical value in percentage of increase in output. For example, a typical figure is 9%, meaning that workers hired with the improved selection method will have on average 9% higher output. A 9% increase in labor productivity would typically be very important economically for the firm, and might make the difference between success and bankruptcy.

What we have presented here is not, of course, a comprehensive discussion of selection utility. Readers who would like more detail are referred to the research articles cited above and to Boudreau (1983a, 1983b, 1984), Cascio and Silbey (1979), Cronshaw and Alexander (1985), Hunter, Schmidt, and Coggin (1988), Hunter and Schmidt (1982, 1983), Schmidt and Hunter (1983), Schmidt, Hunter, Outerbridge, and Trattner (1986), Schmidt, Hunter, and Pearlman (1982), and Schmidt et al. (1984). Our purpose here is to make three important points: (a) the

economic value of gains from improved hiring methods are typically quite large, (b) these gains are directly proportional to the size of the increase in validity when moving from the old to the new selection procedures, and (c) no other characteristic of a personnel measure is as important as predictive validity. If one looks at the two equations above, one sees that practical value per person hired is a three-way product. One of the three elements in that three-way product is predictive validity. The other two—SD_y or SD_p and $\bar{Z}_x$—are equally important, but they are characteristics of the job or the situation, not of the personnel measure.

Validity of Personnel Assessment Methods: 100 Years of Research Findings

Research studies assessing the ability of personnel assessment methods to predict future job performance and future learning (e.g., in training programs) have been conducted since the first decade of the 20th century. However, as early as the 1920s it became apparent that different studies conducted on the same assessment procedure did not appear to agree in their results. Validity estimates for the same method and same job were quite different for different studies. During the 1930s and 1940s the belief developed that this state of affairs resulted from subtle differences between jobs that were difficult or impossible for job analysts and job analysis methodology to detect. That is, researchers concluded that the validity of a given procedure really was different in different settings for what appeared to be basically the same job, and that the conflicting findings in validity studies were just reflecting this fact of reality.

This belief, called the theory of situational specificity of validity, remained dominant in personnel psychology until the late 1970s when it was discovered that most of the differences across studies were due to statistical and measurement artifacts and not to real differences in the jobs (Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman, & Shane, 1979). The largest of these artifacts was simple sampling error variation, caused by the use of small samples in the studies.

(The number of employees per study was usually in the 40–70 range.) This realization led to the development of quantitative techniques collectively called meta-analysis that could combine validity estimates across studies and correct for the effects of these statistical and measurement artifacts (Hunter & Schmidt, 1990; 2004; Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 2015). Studies based on meta-analysis provided more accurate estimates of the average operational validity and showed that the level of real variability of validities was usually quite small and might in fact be zero (Schmidt, 1992; Schmidt et al., 1993). In addition, the findings indicated that the variability of validity was not only small or zero across settings for the same type of job, but was also small across different kinds of jobs of similar complexity (Hunter, 1980; Schmidt, Hunter, & Pearlman, 1981). These findings made it possible to select the most valid personnel measures for any job. They also made it possible to compare the validity of different personnel measures for jobs in general, as we do in this paper.

Table 1 summarizes research findings for the prediction of performance on the job. The third column in Table 1 shows the estimated mean validity of 31 selection methods for predicting performance on the job, as revealed by meta-analyses. Performance on the job was typically measured using supervisory ratings of job performance, but production records, sales records, and other measures were also used. The sources and other information about these validity figures are given in the notes to Table 1.

[Insert Table 1 about here]

Many of the selection methods in Table 1 also predict job-related learning; that is, the acquisition of job knowledge with experience on the job, and the amount learned in training and development programs. However, the overall amount of research on the prediction of learning is less. For many of the procedures in Table 1, there is little research evidence on their ability to

predict future job-related learning. Table 2 summarizes available research findings for the prediction of performance in training programs. The third column in Table 2 shows the mean validity of 16 selection methods as revealed by available meta-analyses. In the vast majority of the studies included in these meta-analyses, performance in training was assessed using objective measures of amount learned on the job; trainer ratings of amount learned were used in about 5% of the studies.

[Insert Table 2 about here]

Unless otherwise noted in Tables 1 and 2, all validity estimates in Tables 1 and 2 are corrected for the downward bias due to measurement error in the measures of job performance and for range restriction on the selection method in incumbent samples relative to applicant populations. No correction is made for measurement error in the predictor scores, because observed scores must be used in selection; true scores are unknown and cannot be used. Observed validity estimates corrected in this manner estimate operational validities of selection methods when used to hire from applicant pools. Operational validities are also referred to as true validities. In both tables selection methods are presented in the order of their incremental validity over that of GMA rather than in the order of their zero order operational validity. In both tables they are numbered from low to high on this dimension and the procedures in Table 1 are discussed in this order.

1. General Mental Ability (GMA)

In the pantheon of 31 personnel measures in Table 1, GMA (also called general cognitive ability and general intelligence) occupies a special place, for several reasons. First, it has the highest validity and lowest application cost. Second, the research evidence for the validity of GMA measures for predicting job performance is stronger than that for any other method

(Hunter, 1986; Hunter & Schmidt, 1996; Ree & Earles, 1992; Schmidt, 2002; Schmidt & Hunter, 1981; Schmidt et al., 2008). Literally thousands of studies of the validity of GMA have been conducted over the last 100 years. By contrast, fewer studies (often far fewer studies) have been conducted on the validity of other selection methods. Third, GMA has been shown to be the best available predictor of job-related learning. It is the best predictor of acquisition of job knowledge on the job (Schmidt & Hunter, 1992; Schmidt, Hunter, & Outerbridge, 1986) and of performance (learning) in job training programs (Hunter, 1986; Hunter & Hunter, 1984; Ree & Earles, 1992; Schmidt et al., 2008). Fourth, the theoretical foundation for GMA is stronger than for any other personnel measure. Theories of intelligence have been developed and tested by psychologists for around 100 years (Brody, 1992; Carroll, 1993; Jensen, 1998). As a result of this massive research literature, the meaning of the construct of intelligence is much clearer than, for example, the meaning of what is measured by other selection procedures such as interviews, situational judgment tests, "emotional intelligence" measures, person-job fit measures, person-organization fit measures, or assessment centers (Arthur & Villado, 2008; Brody, 1992; Hunter, 1986; Jensen, 1998).

The value of .65 in Table 1 for the validity of GMA is the average of eight meta-analytic estimates as presented in Schmidt et al. (2008). Based on the data from a large meta-analytic study conducted for the U.S. Department of Labor (Hunter, 1980; Hunter & Hunter, 1984), Hunter et al. (2006), after applying the new range correction procedure, found that GMA validity ranged from .74 for professional and managerial jobs down to .39 for unskilled jobs. The mean validity for medium complexity jobs (62% of all jobs in the U.S.) was .66. When Schmidt et al. (2008) averaged this value with seven other meta-analytic values the overall average was the .65 seen in Table 1. The medium complexity category includes skilled blue collar jobs and mid-level

white collar jobs, such as upper level clerical and mid to lower level administrative and managerial jobs. Hence, the conclusions in this article apply mainly to the middle 62% of jobs in the U.S. economy in terms of complexity. This figure of .65 produces a high level of practical utility; it produces 65% of the gain in job performance that would be realized with perfectly accurate selection.

As noted above, GMA is also an excellent predictor of job-related learning. It has been found to have high and essentially equal predictive validity for performance (amount learned) in job training programs for jobs at all job levels studied. The average validity value across eight meta-analyses in Schmidt et al. (2008) is .67. This is the figure entered in Table 2. [This is also the average value found in the U.S. Department of Labor research (Hunter et al., 2006).] Thus, when an employer uses GMA to select employees who will have a high level of performance on the job, that employer is also selecting those who will learn the most from job training programs and will acquire job knowledge faster from experience on the job. In fact, most selection procedures shown to be valid for training performance are also valid for predicting job performance, as can be seen by comparing Table 2 to Table 1. Examples include integrity tests, conscientiousness tests, and employment interviews.

Because of its special status, GMA can be considered the primary personnel measure for hiring decisions, and we can consider the remaining 30 personnel measures as supplements to GMA measures. That is, in the case of each of the other measures, one can ask the following question: When used in a properly weighted combination with a GMA measure, how much will each of these measures increase predictive validity for job performance over the .65 that can be obtained by using only GMA? This “incremental validity” translates into incremental utility, that is, into increases in practical value. Because validity is directly proportional to utility, the

percentage increase in validity produced by adding the second measure is also the percentage of increase in practical value (utility). In Tables 1 and 2 selection measures are presented not in the order of their zero order operational validity, but in the order of their incremental contribution to prediction. It will become apparent that many hiring procedures currently receiving a lot of attention in business and other organizations make little or no contribution to predictive validity over that of GMA. In addition, many of these hiring procedures also have low zero order validity.

The increase in validity (and utility) depends not only on the validity of the measure added to GMA, but also on the correlation between the two measures. The smaller this correlation is, the larger is the increase in overall validity. The figures for incremental validity in Table 1 are affected by these correlations. The correlations between mental ability measures and the other measures were estimated from the research literature (often from meta-analyses); the sources of these estimates are given in the notes to Tables 1 and 2. To accurately represent the observed score correlations between predictors in applicant populations, we corrected all correlations between GMA and other predictors for indirect range restriction but not for measurement error in the measure of either predictor.

2. Integrity Tests

Consider integrity tests. These tests are used in business and industry to hire employees with reduced probability of counterproductive work behaviors on the job, such as fighting, drinking or taking drugs, stealing from the employer, equipment sabotage, or excessive absenteeism. Integrity tests do predict these behaviors, but surprisingly they also predict overall job performance (Ones, Viswesvaran, & Schmidt, 1993). The zero order operational validity for job performance is .46. As seen in Table 1, adding an integrity test to the GMA test produces a

.13 increment in validity, a 20% increase in validity and therefore in practical utility. Validity increases from .65 to .78. This largest of incremental of validities is due in significant part to the fact that integrity tests correlate nearly zero with GMA.

A meta-analysis based on 8 studies and 2,364 individuals estimated the mean validity of integrity tests for predicting performance in training programs at .43 (Schmidt, Ones, & Viswesvaran, 1994). As can be seen in Table 2, the incremental validity for integrity tests for predicting training performance is .11, which yields a 16% increase in validity and utility over that produced by GMA alone. In the prediction of training performance, integrity tests appear to produce higher incremental validity than any other measure studied to date. Integrity tests have been found to measure, in part, the personality traits of Conscientiousness, Agreeableness, and Emotional Stability (Ones, 1993). More recent research has shown that integrity tests also measure in part the construct of honesty-humility (Marcus, Lee, & Ashton, 2007). This scale measures tendencies towards being sincere, modest, fair, and honest.

3, 4, & 6. Employment Interviews

Employment interviews can be either structured or unstructured (Huffcutt, Roth, & McDaniel, 1996; McDaniel et al., 1994). Unstructured interviews have no fixed format or set of questions to be answered. In fact, the same interviewer often asks different applicants different questions. Nor is there a fixed procedure for scoring responses; in fact, responses to individual questions are usually not scored, and only an overall evaluation (or rating) is given to each applicant, based on summary impressions and judgments. Structured interviews are exactly the opposite on all counts. In addition, the questions to be asked are usually determined by a careful analysis of the job in question and cannot be deviated from by the interviewer. Structured interviews are more costly to construct and use. Survey results indicate that many employment

interviewers (typically managers) object to having to use structured job interviews because of the restrictions they place on the way they conduct their interviews (Schmidt & Zimmerman, 2004).

Until recently, the available meta-analytic data indicated that the unstructured interview was less valid than the structured interview. Application of the new, more accurate method of correcting for range restriction changed that conclusion (Oh, Postlethwaite, & Schmidt, 2013). As shown in Table 1, the average operational validity of the structured and unstructured interviews is equal at .58. With the former less accurate procedure for correcting for range restriction the validity estimates were .51 for the structured interview and .38 for the unstructured interview. The new figures represent a substantial change, so some explanation is appropriate. The more accurate correction for range restriction requires an estimate of the reliability of the predictor. (See Hunter et al., 2006, for details.) For interviews, the appropriate reliability is estimated as the average correlation between different interviewers interviewing the same group of applicants on different occasions (different days). This form of reliability controls for all three relevant types of measurement error in interviews (Schmidt & Hunter, 2015, pp. 115–121). Two different studies have provided meta-analytic estimates of this reliability for unstructured interviews. Conway, Jako, and Goodman (1995) reported an average reliability of .37, and Huffcutt, Culbertson, and Weyhrauch (2013) obtained an average value of .40. The mean values obtained in these studies are relatively low, causing the range restriction correction to be relatively large. If subsequent meta-analytic estimates of reliability for unstructured interviews are larger, then the estimate of operational validity for the unstructured interview will be lower. [McDaniel et al. (1994) reported higher interview reliabilities but those reliability estimates did not control of all three types of measurement error in interview scores and thus were inflated.]

Although zero order operational validity is about equal for the two types of employment

interviews, the structured interview produces a larger incremental validity—an 18% increase versus a 13% increase for unstructured interviews. This occurs because unstructured interviews have a higher correlation with GMA. As shown in Table 2, unstructured interviews have higher validity than structured interviews for predicting performance in job training programs and also have higher incremental validity (.070 vs. .034). Both these are surprising findings. But the broader conclusion is that employment interviews do predict training performance and do contribute to prediction over and above the validity of GMA measures. Neither of these facts is widely known.

Next, consider phone-administered interviews (6 in Table 1). This is a non-traditional type of structured employment interview in which the questions and keyed answers are developed empirically by comparing the responses of high and low performing employees on the job in question. The interview is conducted by telephone and recorded, with the answers later scored based on the taped transcript. A major advantage of this type of interview is lower costs stemming from the ability to interview geographically scattered applicants, thus saving travel costs. The zero order validity of this interview type is somewhat lower than that for traditional employment interviews but the incremental validity is substantial (.057), producing a 9% increase in validity. This interview type has not been evaluated for its ability to predict training program performance and so does not appear in Table 2.

5. Interests

Earlier research (e.g., Hunter & Hunter, 1984) indicated low validity (.10) for interests in predicting job performance. In the earlier studies these meta-analyses were based on, no attempt was made to match the type of interest measure with the type of job. For example, Realistic interests (in the Holland RIASEC interest model) are relevant to the job of mechanic. The other 5

interest types and their scales are irrelevant to, and inappropriate for use with, the job of mechanic. But in the earlier research the validity of all 6 RIASEC scales was assessed against performance in every kind of job, and as a result mean validities were very low. The validity information for interests reported here is for vocation-focused interest scales that have been matched to the dominant interest classification of job in question. For example, in the case of the mechanics job, only the validity of the Realistic interest scale is included. Other interest scales, such as the Social scale, are not paired with the mechanics job or any of the other jobs classified as Realistic in nature (i.e., as drawing on Realistic interests).

Given this sort of appropriate matching of interest scale to job type, interest measures show an average operational validity of .31 for job performance and produce an incremental validity of .062, or 10%. In Table 2, it can be seen that the situation is similar for the prediction of training program success: an operation validity of .34 and an incremental validity of .070, a validity increase of 11%. Clearly, earlier research created a misleading picture of the potential of interest measures to predict both job performance and amount learned in training programs.

The incremental validity presented in Table 1 is an average across all six RIASEC interest scales. The incremental validity actually varies somewhat across these scales. The incremental validities for the individual scales are provided in note e to Table 1. The reason for the variation in incremental validities is that some interest scales correlate higher with GMA (e.g., Investigative, with $r = .25$) and others correlate lower with GMA (e.g., Artistic, with $r = -.02$). These correlations are also provided in note e.

7. Conscientiousness Measures

The figures for the prediction of job performance from conscientiousness measures are given in Table 1. The validity of conscientiousness measures is substantially lower than that for

integrity tests (.22 vs. .46), its increment to validity is smaller (.053), as is its percentage of increase in validity (8%). However, these values for conscientiousness are still large enough to be practically useful. Turning to Table 2, it can be seen that the validity of conscientiousness for predicting performance in job training programs is .25, producing an incremental validity contribution of .06, a 9% increase in validity. As indicated in the table notes, validity values for all the Big Five personality traits are from the Schmidt et al. (2008) multiple meta-analytic estimates. These values are for standard self-report personality measures.

8. Reference Checks

For the next procedure, reference checks, the incremental validity is 8%, the same value as for Conscientiousness measures. However, the information presented in Table 1 may not at present be fully accurate. The validity studies on which the validity of .26 in Table 1 is based were conducted prior to the development of the current legal climate in the United States. Starting during the 1970s and 1980s, employers providing negative information about past job performance or behavior on the job to prospective new employers were sometimes subjected to lawsuits by the former employees in question. Today, in the United States at least, many previous employers will provide only information on the dates of employment and the job titles the former employee held. Past employers typically refuse to release information on quality or quantity of job performance, disciplinary record of the past employee, or whether the former employee quit voluntarily or was dismissed. This is especially likely to be the case if the information is requested in writing; occasionally, such information will be revealed by telephone or in face to face conversation but one cannot be certain that this will occur.

The legal climate in the United States has changed over the past decades and by the turn of the millennium, 36 states had enacted laws that grant employers immunity from legal liability

for providing good faith job references. Given these changes, one might think that reference checks may again come to provide an increment to the validity of a GMA measure for predicting job performance. In practice, however, reference immunity laws have not had major effects. Employers are still reluctant to provide more than dates of employment and job titles (Cooper, 2001, p. 14).

Older research indicates that reference checks predict performance in training with a mean validity of .23 (Hunter & Hunter, 1984, Table 8), yielding a 6% increment in validity over GMA tests, as shown in Table 2. But, again, these findings may no longer hold; it seems unlikely that changes in the legal climate will make these validity estimates accurate again.

9. Openness to Experience Measures

Measures of the personality trait of Openness to Experience have very low zero order validity (.04) but nevertheless produces an increase in overall validity of 6%. This occurs because Openness to Experience functions as a suppressor variable, as can be seen from its negative regression weight in Table 1. Use of that negative regression weight is required for attainment of the listed incremental validity. Despite the 6% incremental validity produced by the suppressor effect, most employers would probably be reluctant to use a predictor with a zero order operational validity as low as .04. Employers would also probably be reluctant to employ the negative regression weight. The operation and function of suppressor variables is explained in detail in Collins and Schmidt (1997). As can be seen in Table 2, Openness to Experience has no incremental validity for prediction of training performance, despite the fact that its zero order operational validity (.24) is higher than it is for predicting job performance (.04). Among the Big Five personality traits, Openness to Experience has the largest correlation with GMA (.38), which limits its incremental validity. (These correlations are provided in note g to Table 1.)

10. Biographical Data Measures (Biodata)

Next in Table 1 is biographical data measures, usually referred to as biodata. Biographical data measures contain questions about past life experiences, such as early life experiences in one's family, in high school, and in hobbies and other pursuits. For example, there may be questions on offices held in student organizations, on sports one participated in, and on disciplinary practices of one's parents. Each question has been chosen for inclusion in the measure because in the initial developmental sample it correlated with a criterion of job performance, performance in training, or some other criterion (e.g., promotion rate). That is, biographical data measures are empirically developed. However, they are usually not completely actuarial, because some hypotheses are invoked in choosing the beginning set of items. However, choice of the final questions to retain for the scale is mostly actuarial. Today antidiscrimination laws prevent certain questions from being used, such as sex, marital status, and age, and such items are not included. Biographical data measures have been used to predict performance on a wide variety of jobs, ranging in level from blue collar unskilled jobs to scientific and managerial jobs. These measures are also used to predict job tenure (turnover) and absenteeism (cf. Schmidt & Hoffman, 1973), but we do not consider these usages in this article.

Table 1 shows that biographical data measures have substantial zero-order validity (.35) for predicting job performance and produce an increment in validity over GMA of .036 on average (a 6% increase). The reason that the increment in validity is not larger is that biographical data correlates substantially with GMA (Schmidt, 1988). This suggests that in addition to whatever other traits they measure, biographical data measures are also in part indirect reflections of mental ability. As shown in Table 2, biographical data measures predict performance in training programs with a mean validity of .30 and increment overall validity by

.07, an 11% increase. In the prediction of both job performance and training performance, biographical data measures function as suppressor variables, as can be seen from their negative regression weights. Attainment of the incremental validities shown in the tables requires use of these negative regression weights.

Biographical data measures are technically difficult and time consuming to construct (although they are easy to use once constructed). Considerable statistical sophistication and large data sets are required to develop them. However, some commercial firms offer validated biographical data measures for particular jobs (e.g., first line supervisors, managers, clerical workers, and law enforcement personnel). These firms maintain control of the proprietary scoring keys and the scoring of applicant responses.

11. Job Experience (Years)

Job experience as indexed in Tables 1 and 2 refers to the number of years of previous experience on the same or similar job; it conveys no information on past performance on the job. In the data used to derive the validity estimates in these tables, job experience varied widely: from less than 6 months to more than 30 years. Under these circumstances, the validity of job experience for predicting future job performance is only .16 and the increment in validity (and utility) over that from GMA alone is only .032 (a 5% increase). However, Schmidt, Hunter, and Outerbridge (1986) found that in groups in which experience on the job does not exceed 5 years, the correlation between amount of job experience and job performance is considerably larger: .33 when job performance is measured by supervisory ratings, and .47 when job performance is measured using a work sample test. These researchers found that the relation is nonlinear: Up to about 5 years of job experience, job performance increases linearly with increasing experience on the job. After that, the curve becomes increasingly horizontal, and further increases in job

experience produce little increase in job performance. Apparently, during the first 5 years on these (mid-level, medium complexity) jobs, employees were continually acquiring additional job knowledge and skills that improved their job performance. But by the end of 5 years this process was nearly complete, and further increases in job experience led to little increase in job knowledge and skills (Schmidt & Hunter, 1992). These findings suggest that even under ideal circumstances, job experience at the start of median complexity jobs will predict job performance only for the first 5 years on the job. (This period may be longer for professional, scientific, and high level managerial jobs, but this has not been studied.) By contrast, GMA continues to predict job performance indefinitely (Hunter & Schmidt, 1996; Schmidt, Hunter, Outerbridge, & Goff, 1988; Schmidt, Hunter, Outerbridge, & Trattner, 1986).

As shown in Table 2, the amount of job experience does not predict performance in training programs teaching new skills. However, one can note from this finding that job experience does not retard the acquisition of new job skills in training programs as might have been predicted from theories of proactive inhibition.

12 & 22. "Emotional Intelligence" Measures

In the last several decades, "emotional intelligence" measures have become popular in many businesses, and emotional intelligence has been treated in the popular press as an actual psychological trait or construct. We have placed the label emotional intelligence in quotation marks because most psychologists, including personnel psychologists, do not accept that emotional intelligence is a trait or construct; rather they view emotional intelligence measures as arbitrary amalgamations of items measuring long established psychological traits (cf. Murphy, 2006; Matthews, Zeidner, & Roberts, 2002). There are of two types of emotional intelligence measures: personality-based measures (12 in Table 1) and ability based measures (Mayer,

Salovey, Caruso, & Sitarenios, 2003; 22 in Table 1). The questions on ability based measures have answers viewed as correct and these measures are correlated about .45 with GMA when corrected for indirect range restriction (Joseph & Newman, 2010). Hence, they are indicators of GMA, just as verbal and quantitative ability tests are (cf. Schmidt, 2011). Personality-based measures are made up mostly of personality type questions with some other types on non-cognitive measures being included. The zero order validity of personality-based EI measures is .32, with incremental validity of .029, a 5% increase. Ability-based EI measures have an average zero order validity of .23, with basically no incremental validity.

Both the validity and the incremental validity are lower than assumed by advocates of these measures. Emotional intelligence measures have not been studied in the prediction of training performance and so do not appear in Table 2.

13. Person-Organization Fit Measures

Person-Organization Fit measures assess the degree of match between characteristics of the applicants (such as values, goals, desires, and interests) and the values, purposes, and goals of the organization as a whole. These measured do not include any cognitive, ability, or skills component. Measures of Person-Organization Fit have recently become popular in business and industry. For the prediction of job performance they have a low average validity (.13) and produce an incremental validity increase of only 4%. Both these figures are disappointing to the advocates of these measures. Person-organization fit measures have not been studied in relation to performance in job training programs and so do not appear in Table 2.

14. Knowledge-Based Situational Judgment Tests

Next we have knowledge-based Situational Judgment Tests. These measures present the applicant with complex situational workplace problems, often involving human interactions, with

the applicant being required to select the best solution from those listed. These measures have an average validity of .26. However, because they correlate nearly .60 with GMA, their incremental validity is only 2% (.015). Note that these measures function as suppressor variables, with a regression weight of -.17. Attainment of the small incremental validity that they produce requires use of this negative regression weight. No meta-analyses relating these measures to training performance have been conducted, so there is no entry for them in Table 2.

15. Person-Job Fit Measures

The next predictor is Person-Job Fit measures. These measures assess the degree of match between characteristics of the applicants (such as values, desires, and interests) and those embodied in, or offered by, the job the applicant is applying for. They do not include matching on applicant GMA and job GMA requirements. The process of constructing these fit measures for each job can be time consuming and costly. These measures have an average validity of .18 and produce only a 2% increment in validity over that of GMA. Both the zero order operational validity and the incremental validity are lower than had been anticipated by those researching these measures. No studies have been conducted on the validity of Person-Job Fit measures for predicting training performance and so they do not appear in Table 2.

16. Assessment Centers

Assessment centers are expensive to use. Individuals who are administered assessment centers spend one to several days at a central location where they are observed participating in such exercises as leaderless group discussions, business games, and in-basket exercises. Various ability and personality tests are usually administered, and in-depth structured interviews are also part of most assessment centers. The average assessment center includes seven exercises or assessments and lasts 2 days (Gaugler, Rosenthal, Thornton, & Benson, 1987). Assessment

centers are used for jobs ranging from first line supervisors to high level management positions. The observers who record and score the performance of the candidates are personnel psychologists or manager trained in the observer role by personnel psychologists.

Assessment centers are like biographical data measures: They have substantial validity (.36) but little incremental validity over GMA (.01, a 2% increase) in predicting job performance. The reason is also the same: They correlate highly with GMA—in part because they typically include a measure of GMA (Gaugler et al., 1987; Collins et al., 2003). Assessment center scores function as suppressor variables when used with GMA. Despite the fact of relatively low incremental validity, many organizations use assessment centers for managerial jobs because they believe assessment centers provide them with a wide range of insights about candidates and their developmental possibilities. As can be seen in Table 2, assessment centers show a similar level of validity for the prediction of performance in job training programs: .37, with an incremental validity of 2%.

17. The Training and Experience Point Method

The point method of evaluating previous training and experience (T & E) is used mostly in government hiring at all levels—federal, state, and local. A major reason for its widespread use is that point method procedures are relatively inexpensive to construct and use. The point method appears under a wide variety of different names (McDaniel et al., 1988), but all such procedures have several important characteristics in common. All point method procedures are credentialistic; typically, an applicant receives a fixed number of points for (a) each year or month of experience on the same or similar job, (b) each year of schooling or each course taken that appears to be face-relevant, and (c) each relevant training program completed, and so on. There is usually no attempt to evaluate past achievements, accomplishments, or job performance;

in effect, the procedure assumes that achievement and performance are determined solely by the exposures that the method measures. As shown in Table 1, the T & E point method has low validity (.11) and produces only a 1% increase in validity over that available from GMA alone. The T & E point method has not been used to predict performance in training programs and so does not appear in Table 2.

18. Grade Point Average

The validity value for grade point average (GPA) in Table 1 is for college and graduate level grade point averages. No estimates are available for high school grade point average, which may have validity higher than the .34 in Table 1. Apparently most of the validity of GPA is captured by GMA, because the incremental validity of GPA is negligible (less than .01). GPA has not been studied in relation to training performance, where its validity might be expected to be higher than the .34 for job performance, because of the strong resemblance between training programs and classroom demands.

19. Years of Education

Sheer amount of education has even lower validity for predicting job performance than the T & E point method (.10). Its increment to validity is the same 1% as obtained with the T & E point method. It is important to note that this finding does not imply that education is irrelevant to occupational success; education is clearly an important determinant of the level of job the individual can obtain. What this finding shows is that among those who apply to get a particular job, years of education does not predict future performance on that job very well. For example, for a typical blue collar job, years of education among applicants might range from 9 to 12. The validity of .10 then means that the average job performance of those with 12 years of education will be only very slightly higher (on average) than that for those with 9 or 10 years.

As can be seen in Table 2, amount of education predicts learning in job training programs better than it predicts performance on the job. Hunter and Hunter (1984, Table 6) found a mean validity of .20 for performance in training programs. This is not a high level of validity, but it is twice as large as the validity for predicting job performance. Its incremental validity is .03 (a 4% increase).

20. Extraversion

As can be seen in Table 1, measures of the personality trait of Extroversion have low validity for predicting job performance (.09) and close to no incremental validity. These measures perform somewhat better in the prediction of training performance (.17), with an incremental validity of .02 (3%). Refer to note g in Table 1 for additional information.

21. Peer Ratings

Peer ratings are evaluations of performance or potential made by one's co-workers; they typically are averaged across peer raters to increase the reliability (and hence validity) of the ratings. Peer ratings have the limitation that they cannot be used for evaluating and hiring applicants from outside the organization; they can be used only for internal job assignment, promotion, or training assignment. They have been used extensively for these internal personnel decisions in the military (particularly the U.S. and Israeli militaries) and some private firms, such as insurance companies. One concern associated with peer ratings is that they will be influenced by friendship, or social popularity, or both. Another is that pairs or clusters of peers might secretly agree in advance to give each other high peer ratings. However, the research that has been done does not support these fears; for example, partialling friendship measures out of the peer ratings does not appear to affect the validity of the ratings (cf. Hollander, 1956; Waters & Waters, 1970).

As shown in Table 1, despite having a relatively high zero order operational validity (.49), peer ratings produce only a 1% increase in validity over that of GMA in the prediction of job performance. This occurs because peer ratings correlate highly with GMA (nearly .60). As seen in Table 2, validity for training performance is .36 but incremental validity is only 1%, again due to the high correlation with GMA. Peer ratings are the last predictor in Table 1 with incremental validity or 1% or larger. The remainder are at less than 1%. This includes some that are widely used and highly regarded.

22. Next is Ability-Based measures of "emotional intelligence". The discussion of these measures is presented along with that of Personality-Based EI measures (12, above). Ability-Based EI measures have a mean operational validity of .23 and produce no incremental validity. The validity of these measures for the prediction of learning in job training programs has not been studied, and so there is no entry for them in Table 2.

23 & 26. Measures of Agreeableness and Emotional Stability

Next, consider measures of the Big Five personality traits of Agreeableness and Emotional Stability. In Table 1 we can see that both these trait measures have low validity for predicting job performance and no incremental validity. In Table 2, it can be seen that their validities are also low for training performance, and incremental validity is zero for Emotional Stability and only 1% for Agreeableness. The reader should keep in mind that these values are for standard self-report measures; personality traits measured using ratings by others who know the individual (e.g., coworkers) tend to have higher validities (Oh, Wang, & Mount, 2011). However, such measures are rarely used in job selection at present because such ratings are difficult to obtain. It is also the case that validities are somewhat higher when the questions asked are made work-specific; that is, they ask how the respondent behaves or reacts at work or

on the job, rather than in general (Shaffer & Postlethwaite, 2012, Table 1). This approach to increasing validity is not difficult to implement. Refer to note g in Table 1 for additional information.

24. Work Sample Tests

Next, consider work sample tests. Work sample tests are hands-on simulations of part or all of the job that must be performed by applicants. For example, as part of a work sample test, an applicant might be required to repair a series of defective electric motors or computers. Work sample tests are often used to hire skilled workers, such as welders, machinists, and carpenters. These tests can be used only with applicants who already know the job or occupation. Such workers do not need to be trained, and so the ability of work sample tests to predict training performance has not been studied. Hence, there is no entry for work sample tests in Table 2.

In the earlier Schmidt and Hunter (1998) article, the average validity of work sample tests for predicting job performance was reported as .54, based on the meta-analysis of Hunter and Hunter (1984). At that time there were only seven primary studies available to be included in their meta-analysis. Since 1984 many more studies have been conducted; these studies were added to the earlier studies and meta-analyzed by Roth, Bobko, and McFarland (2005), producing the mean validity of .33 presented in Table 1. Roth et al. (2005) attributed the decrease in the estimated validity of work sample tests produced by the newer studies to increased use of work sample tests in the service sector of the economy. It is possible that the original validity estimate of .54 is still applicable to traditional manual skilled trades such as machinist, carpenter, welders, and the like.

It appears that almost all of the validity of work sample tests is captured by GMA measures, because the incremental validity is essentially zero. This case, like that of unstructured

interviews (discussed earlier), is an example of the fact that in science new research can change results and estimates.

25. Situational Judgment Tests (Behavioral Tendency). Refer to 12, above.

26. Emotional Stability measures. Refer to 23, above.

27. Graphology

Graphology is the analysis of handwriting. Graphologists claim that people express their personalities through their handwriting and that one's handwriting therefore reveals personality traits and tendencies that graphologists can use to predict future job performance. Graphology is used infrequently in the United States and Canada but is widely used in hiring in France (Steiner, 1997; Steiner & Gilliland, 1996), although not in most other European countries (Bangerter, Konig, Blatti, & Salvisberg (2009). Levy (1979) reported that 85% of French firms routinely use graphology in hiring of personnel. Ben-Shakhar, Bar-Hillel, Bilu, Ben-Abba, and Flug (1986) stated that in Israel graphology is used more widely than any other single personality measure.

Several studies have examined the ability of graphologists and nongraphologists to predict job performance from handwriting samples (Jansen, 1973; Rafaeli & Klimoski, 1983; see also Ben-Shakhar, 1989; Ben-Shakhar, Bar-Hillel, Bilu, et al., 1986; Ben-Shakhar, Bar-Hillel, & Flug, 1986). The key findings in this area are as follows. When the assessees who provide handwriting samples are allowed to write on any subject they choose, both graphologists and untrained non-graphologists can infer some (limited) information about their personalities and job performance from the handwriting samples. But untrained non-graphologists do just as well as graphologists; both show validities in the .18–.20 range. When the assesses are required to copy the same material from a book to create their handwriting sample, the evidence indicates that neither graphologists nor non-graphologists can infer any valid information about

personality traits or job performance from the handwriting samples (Neter & Ben-Shakhar, 1989).

What these findings indicate is that, contrary to graphology theory, whatever limited information about personality or job performance there is in the handwriting samples comes from the content and not the characteristics of the handwriting. For example, writers differ in style of writing, range of vocabulary, expression of emotions, verbal fluency, grammatical skills, and general knowledge. Whatever information about personality and ability these differences contain, the training of graphologists does not allow them to extract it better than can people untrained in graphology. In handwriting per se, independent of content, there appears to be no information about personality or job performance (Neter & Ben-Shakhar, 1989).

To many people, this is a counterintuitive finding. To these people, it seems obvious that the wide and dramatic variations in handwriting that everyone observes must reveal personality differences among individuals. Actually, most of the variation in handwriting is due to differences among individuals in fine motor coordination of finger muscles. And these differences in finger muscles and their coordination are probably due mostly to random genetic variations among individuals. The genetic variations that cause these finger coordination differences do not appear to be linked to personality; and in fact there is no apparent reason to believe they should be.

The validity of graphology for predicting performance in training programs has not been studied. However, the findings with respect to performance on the job make it highly unlikely that graphology has validity for training performance.

28. Job Tryout Procedure (Internship)

Next, consider the job tryout procedure. Unlike work sample tests, the job tryout

procedure can be used with entry level employees with no previous experience on the job in question. With this procedure, applicants are hired with minimal or no screening and their performance on the job is observed and evaluated for a certain period of time (typically 6–8 months). Those who do not meet a previously established standard of satisfactory performance by the end of this probationary period are then terminated. If used in this manner, this procedure can have substantial validity (.44), as shown in Table 1. However, it is very expensive to implement, and low job performance by minimally screened probationary workers can lead to serious economic losses. In addition, it has been our experience that supervisors are reluctant to terminate marginal performers. Doing so is an unpleasant experience for them, and to avoid this experience many supervisors gradually reduce the standards of minimally acceptable performance, thus diluting the effectiveness of the procedure. Another consideration is that some of the benefits of this method will be captured in the normal course of events even if the job tryout procedure is not used, because clearly inadequate performers will usually be terminated after a period of time anyway. Also, the job tryout procedure has no incremental validity over that of GMA, because of its high correlation with GMA (.66). If this procedure were used in connection with training programs, it would probably be called the Training Tryout Procedure. However, we know of no cases in which this has been done.

29. The Behavior Consistency Method.

The behavioral consistency method of evaluating previous training and experience (McDaniel, Schmidt, & Hunter, 1988; Schmidt, Caplan, et al., 1979) is based on the well-established psychological principle that the best predictor of future performance is past performance. In developing this method, the first step is to determine what achievement and accomplishment dimensions best separate top job performers from low performers. This is done

on the basis of information obtained from experienced supervisors of the job in question, using a special set of procedures (Schmidt, Caplan, et al., 1979). Applicants are then asked to describe (in writing or sometimes orally) their past achievements that best illustrate their ability to perform these functions at a high level (e.g., organizing people and getting work done through people). These achievements are then scored with the aid of scales that are anchored at various points by specific scaled achievements that serve as illustrative examples or anchors

Use of the behavioral consistency method is not limited to applicants with previous experience on the job in question. Previous experience on jobs that are similar to the current job in only very general ways typically provides adequate opportunity for demonstration of achievements. In fact, the relevant achievements can sometimes be demonstrated through community, school, political, and other nonjob activities. However, some young people just leaving secondary school may not have had adequate opportunity to demonstrate their capacity for the relevant achievements and accomplishments; the procedure might work less well in such groups.

A behavioral consistency procedure can be time consuming and costly to create. Considerable work is required to construct the procedure and the scoring system; applying the scoring procedure to applicant responses is also more time consuming than scoring of most tests with clear right and wrong answers. However, especially for higher level jobs, the behavioral consistency method may be well worth the cost and effort in situations in which it is not possible to use a GMA test. As shown in Table 1, this procedure has a validity of .45 for predicting job performance but produces no incremental validity over that of GMA, due to its substantial correlation with GMA. No information is available on the validity the behavioral consistency procedure for predicting performance in training programs.

30. Job Knowledge Tests

The next procedure in Table 1 is job knowledge tests. Like work sample measures, job knowledge tests cannot be used to evaluate and hire inexperienced workers. An applicant cannot be expected to have mastered the job knowledge required to perform a particular job unless he or she has previously performed that job or has received schooling, education, or training for that job. But applicants for jobs such as carpenter, welder, accountant, and chemist can be administered job knowledge tests. Job knowledge tests are often constructed by the hiring organization on the basis of an analysis of the tasks that make up the job. Constructing job knowledge tests in this manner is generally somewhat more time consuming and expensive than constructing typical structured interviews. However, such tests can also be purchased commercially; for example, tests are available that measure the job knowledge required of machinists (knowledge of metal cutting tools and procedures). Other examples are tests of knowledge of basic organic chemistry and tests of the knowledge required of roofers. In an extensive meta-analysis, Dye, Reck and McDaniel (1993) found that commercially purchased job knowledge tests ("off the shelf" tests) had slightly lower validity than job knowledge tests tailored to the job in question. The validity figure of .48 in Table 1 for job knowledge tests is for tests tailored to the job in question.

As shown in Table 1, job knowledge tests do not increment validity over that of GMA, because their high correlations with GMA. For the same reasons indicated earlier for job sample tests, job knowledge tests typically have not been used to predict performance in training programs. Hence, validity information is unavailable for this criterion, and there is no entry in Table 2 for job knowledge tests.

31. Age

Table 1 shows that age of job applicants shows no validity for predicting job performance. Age is rarely used as a basis for hiring, and in fact in the United States, use of age for individuals over age 40 would be a violation of the federal law against age discrimination. We include age here for only two reasons. First, some individuals believe age is related to job performance. We show here that for typical jobs this is not the case. Second, age serves to anchor the bottom end of the validity dimension: Age is about as totally unrelated to job performance as any measure can be. No meta-analyses relating age to performance in job training programs were found. Although it is possible that future research will find that age is negatively related to performance in job training programs (as is widely believed), we note that job experience, which is positively correlated with age, is not correlated with performance in training programs (see Table 2).

An Important Question

Finally, we address an issue that some in our field have raised (e.g., Arthur & Villado, 2008). As discussed in more detail in the next section, some of the personnel measures we have examined (e.g., GMA and conscientiousness measures) are measures of single psychological constructs, whereas others (e.g., biodata and assessment centers) are methods rather than constructs. It is conceivable that a method such as the assessment center, for example, could measure different constructs or combinations of constructs in different applications in different organizations. Some would, therefore, question whether it is meaningful to compare the incremental validities of different methods (e.g., comparing the incremental validities produced by the structured interview and the assessment center). There are two responses to this. First, this article is concerned with personnel measures as used in the real world of employment. Hence, from that point of view, such comparisons of incremental validities would be meaningful, even if

they represented only crude average differences in incremental validities.

However, the situation is not that grim. The empirical evidence indicates that such methods as interviews, assessment centers, and biodata measures do not vary much from application to application in the constructs they measure. This can be seen from the fact that meta-analysis results show that variability of validity across studies (applications) of different jobs, after the appropriate corrections for sampling error and other statistical and measurement artifacts, is quite small (cf. Gaugler et al., 1987; McDaniel et al., 1994; Schmidt & Rothstein, 1994). In fact, these standard deviations are often even smaller than those for construct-based measures such as GMA and conscientiousness (Schmidt & Rothstein, 1994). Hence, the situation appears to be this: We do not know exactly what combination of constructs is measured by methods such as the assessment center, the interview, and biodata, but whatever those combinations are, they do not appear to vary much from one application or study to another. If they did vary, one would expect the resulting validities to vary but they don't. Hence, comparisons of their relative incremental validities over GMA is in fact meaningful. These incremental validities can be expected to be stable across different applications of the methods to different jobs in different organizations and settings.

Toward a Theory of the Determinants of Job Performance

Earlier sections in this paper summarized what is known from cumulative empirical research about the validity of various personnel measures for predicting future job performance and job-related learning of job applicants. These findings are based on thousands of research studies performed over the last century and involving millions of employees and job applicants. They are a tribute to the power of empirical research, integrated using meta-analysis methods, to produce precise estimates of relationships of interest and practical value. However, the goals of

personnel psychology include more than a delineation of relationships that are practically useful in selecting employees. There is also a focus on development of theories of the causes of job performance (Schmidt & Hunter, 1992). The objective is the understanding of the psychological processes underlying and determining job performance. This endeavor is possible because application of meta-analysis to research findings has provided the kind of precise and generalizable estimates of the validity of different measured constructs for predicting job performance that are summarized in this paper. It has also provided more precise estimates than previously available of the correlations among these predictors.

However, the theories of job performance that have been developed and tested do not include a role for all of the personnel measures discussed above. That is because the actual constructs measured by some of these procedures are unknown, and it seems certain that some of these procedures measure combinations of constructs (Hunter & Hunter, 1984; Schmidt & Rothstein, 1994). For example, employment interviews probably measure a combination of previous experience, mental ability, and a number of personality traits, such as conscientiousness; in addition, they may measure specific job-related skills and behavior patterns. The average correlation between scores on unstructured interviews and scores on GMA tests is .31 (see note c to Table 1). This indicates that, to some extent, interview scores reflect mental ability. Little empirical evidence is available as to what other traits they measure (Huffcutt et al., 1996). What has been said here of employment interviews also applies to peer ratings, the behavioral consistency method, reference checks, biographical data measures, assessment centers, and the point method of evaluating past training and experience. Procedures such as these can be used as practical selection tools but, because their construct composition is unknown, they are less useful in constructing theories of the determinants of job performance.

The measures that have been used in theories of job performance have been GMA, job knowledge, job experience, and personality traits. This is because it is fairly clear what constructs each of these procedures measures.

What has this research revealed about the determinants of job performance? A detailed review of this research can be found in Schmidt and Hunter (1992); here we summarize only the most important findings. One major finding concerns the reason why GMA is such a good predictor of job performance. The major direct causal impact of GMA has been found to be on the acquisition of job knowledge. That is, the major reason more intelligent people have higher job performance is that they acquire job knowledge more rapidly and acquire more of it; and it is this knowledge of how to perform the job that causes their job performance to be higher (Hunter, 1986). Thus, GMA has its most important effect on job performance indirectly, through job knowledge. There is also a direct effect of mental ability on job performance independent of job knowledge, but it is smaller for most jobs. For nonsupervisory jobs, this direct effect is only about 20% as large as the indirect effect; however, for supervisory jobs, it is about 50% as large (Borman, White, Pulakos, & Oppler, 1991; Schmidt, Hunter, & Outerbridge, 1986).

It has also been found that job experience operates in this same manner. Job experience is essentially a measure of practice on the job and hence a measure of opportunity to learn. The major direct causal effect of job experience is on job knowledge, just as is the case for mental ability. Up to about 5 years on the job, increasing job experience leads to increasing job knowledge (Schmidt, Hunter, & Outerbridge, 1986), which, in turn, leads to improved job performance. So the major effect of job experience on job performance is indirect, operating through job knowledge. Again, there is also a direct effect of job experience on job performance, but it is smaller than the indirect effect through job knowledge (about 30% as large).

The major personality trait that has been studied in causal models of job performance is conscientiousness. This research has found that, controlling for mental ability, employees who are higher in conscientiousness develop higher levels of job knowledge, probably because highly conscientious individuals exert greater efforts and spend more time "on task." This job knowledge, in turn, causes higher levels of job performance. From a theoretical point of view, this research suggests that the central determining variables in job performance are GMA, job experience (i.e., opportunity to learn), and the personality trait of conscientiousness. This is consistent with our conclusion that a combination of a GMA test and an integrity test (which measures mostly conscientiousness) has the highest high validity (.78) for predicting job performance. Another combination with high validity (.76) is GMA plus a structured interview, which may in part measure conscientiousness and related personality traits (such as agreeableness and emotional stability, which are also measured in part by integrity tests)

Additional Considerations

When GMA Measures cannot be used

There are some situations in which it may not be feasible to use measures of GMA in hiring. For example, higher level managers applying for jobs may insist on being evaluated based on their past achievements. They may readily accept employment interviews, the Behavioral Consistency Method, work sample tests (such as the In-Basket test), or an assessment center, but balk at test of general intelligence, viewing it as less relevant. This may also be true of difficult-to-recruit skilled tradespeople, such as skilled machine repairmen, computer programmers, machinists, and the like. In situations like this, employers should focus on the zero order operational validities in Tables 1 and 2 rather than the incremental validity values. Non-GMA procedures with high zero order validities include employment interviews, job knowledge tests, the Behavioral Consistency

Method, and work sample tests.

Selection Fairness and Adverse Impact

A full treatment of questions related to gender or minority subgroups are beyond the scope of this study. These issues include questions of differential validity by subgroups, predictive fairness for subgroups, and subgroup differences in mean score on selection procedures. An extensive existing literature addresses these questions (cf. Hunter & Schmidt, 1996; Ones et al., 1993; Schmidt, 1988; Schmidt & Hunter, 1981; Schmidt, Ones, & Hunter, 1992; Wigdor & Garner, 1982). However, the general findings of this research literature can be summarized here.

For differential validity, the general finding has been that validities (the focus of this study) do not differ appreciably for different subgroups (Hartigan & Wigdor, 1989; Hunter, Schmidt, & Hunter, 1979; Rothstein & McDaniel, 1992; Schmidt, Berner, & Hunter, 1973). For predictive fairness, the usual finding has been a lack of predictive bias for minorities and women (Schmidt & Hunter, 1981). That is, given similar scores on selection procedures, later job performance is similar regardless of group membership and regardless of how job performance is measured (objectively or via supervisor ratings). On some selection procedures (in particular, cognitive measures), subgroup differences on means are typically observed. On other selection procedures (in particular, personality and integrity measures), subgroup differences are rare or nonexistent. For many selection methods (e.g., reference checks and evaluations of education and experience), there is little data on group differences (Hunter & Hunter, 1984). For many purposes, the most relevant finding is the finding of lack of predictive bias. That is, even when subgroups differ in mean score, selection procedure scores appear to have the same implications for later performance for individuals in all subgroups (Wigdor & Garner, 1982; Schmidt &

Hunter, 1981). That is, the predictive interpretation of scores is the same in different subgroups.

Despite the research demonstration of predictive fairness for selection methods, there are often mean score differences between different demographic groups, such as black and white job applicants, leading to differential hiring rates. When hiring rates are lower for minority or female applicants, this is referred to as "adverse impact" in the context of U.S. legal regulations. (These regulations are absent or different in other countries.) There are often large enough black-white differences in GMA test scores to produce differential hiring rates and hence adverse impact. In terms of sex differences, it has been found that there is no difference in mean levels of GMA between males and females (Ployhart & Holtz, 2008; Schmidt, 2011), although there are some well-known sex differences in specific aptitudes (i.e., higher mean scores for females in verbal fluency and higher mean scores for males in spatial rotation). Though we mention differences in specific aptitudes, we do not recommend the use of specific abilities in staffing contexts, given that the validity of GMA is higher than that of specific aptitudes—even when specific aptitudes are chosen to match the most important aspects of job performance (i.e., spatial perception for mechanical jobs; cf. Schmidt, 2011). The research surrounding questions of ethnic group differences in test scores is quite extensive. We refer readers who are interested in a more technical, detailed treatment of this question to Roth, Bevier, Bobko, Switzer, and Tyler (2001); Sackett, Schmitt, Ellingson, and Kabin (2001); Sackett, Borneman, and Connelly (2008); Schmidt (1988); and Wigdor and Garner (1982).

Lower hiring rates for minority applicants can lead to two potential problems: Legal risks and decreased workplace diversity. From a legal standpoint, GMA assessments are quite defensible in court via a demonstration that the tests are valid predictors of job performance. Such demonstrations rely are increasingly based on summaries of the kinds of research findings

discussed in this paper, rather than on studies conducted by the employer. This is part of the movement away from the false notion of situationally specific validity, as discussed earlier. Since around the mid-1980s, employers have been winning more and more such suits, and today they prevail in 90% or more of such suits. A key fact is that today there are far fewer such suits. Currently, less than 1% of employment-related lawsuits are challenges to selection tests or other hiring procedures, and from a purely economic standpoint, research shows that the value of the increases in job performance from good selection practices overshadows any potential costs stemming from defending against such suits. Thus, there is little legal risk stemming from the use of GMA assessments.

However, solving the problem of legal defensibility still leaves the potential problem of workplace diversity, given that many hiring managers seek to increase the diversity of their workforce. Sackett and colleagues (2001) explain that emphasizing GMA assessments in selection with the purpose of maximizing job performance typically results in a lower ratio of minority hires, while eschewing the use of GMA assessments with the purpose of increasing workplace diversity typically results in a decrease in job performance. These authors go on to ask, "What are the prospects for achieving diversity without sacrificing the predictive accuracy and content relevancy present in knowledge, skill, ability, and achievement tests?" (p. 303). Over the years, employers have tried to reduce differences in hiring rates through a variety of strategies—adjusting test scores to minimize between-group differences in scores, assigning more weight to predictors associated with less adverse impact, and using more non-cognitive selection methods—but none of these strategies have been really effective (Ployhart & Holtz, 2008); also, some of the strategies, such as score adjustments, are illegal in the U.S. Along these lines, Pyburn, Ployhart, and Kravitz write, "…an organization might use a less valid selection procedure simply because it has less adverse

impact. Doing so violates no laws, but it fails to capitalize on over 80 years of research that has shown valid selection procedures can enhance job performance and utility" (2008, p. 150). Some have advocated as a solution to this problem the use of GMA assessments supplemented with the use of non-cognitive predictors such as structured interviews, integrity tests, or measures of conscientiousness that show little, if any, group differences. The idea is that a composite battery of predictors could be assembled that not only will be an excellent predictor of job performance, but also will help reduce group differences associated with GMA measures. However, Sacket et al. (2001) showed that this approach produces only very small reductions in differential hiring rates.

Applicant reactions

Consider the following question: Can the use of valid assessment methods during the hiring process be a detriment to a firm's recruitment efforts because job applicants do not like the methods used? This question might be most relevant to tests of GMA since it is the most valid predictor of job and training performance. Fortunately for hiring managers, research on this subject reveals that applicants perceive mental ability and GMA tests to be relevant to job performance. Applicants generally react more favorably to GMA tests than they do to personality tests, biodata, and integrity tests, which they view as less relevant to job performance (Anderson, Salgado, & Hülsheger, 2010; Hausknecht, Day, & Thomas, 2004). Anderson et al. (2010) found that applicants were most favorable towards work samples and interviews, but they also had a favorable view of GMA tests. In fact, although GMA tests were not applicants' most preferred selection method, they viewed GMA as being the most scientifically valid and as being the most respectful of their personal privacy (Anderson, et al., 2010).

Interestingly, Anderson et al. (2010) found that when GMA or other mental ability tests are used, applicants perceive the selection requirements as being more stringent, which increases their

perceptions of the status and attractiveness of the job in question. The conclusion is this: Job applicants generally respond well to GMA tests and other selection methods when they believe they are relevant to job performance, and there is strong correlation between the validity of selection method and favorable applicant reactions to the methods (Anderson et al, 2010).

Limitations of This Study

This article examined the multivariate validity of only certain predictor combinations: combinations of two predictors with one of the two being GMA. Organizations sometimes use more than two selection methods. For example, when hiring an entry level manager an organization might combine the GMA score, interview score, Conscientiousness score and college GPA into some sort of composite score. Therefore, it might be informative to examine the incremental validity resulting from adding these three predictors to GMA. For some purposes, it would also be of interest to examine predictor combinations that do not include GMA. However, the absence of the needed estimates of predictor inter-correlations in the literature makes this impossible at the present time. In the future, as data accumulates, such analyses may become feasible. However, based on the results reported in this study, it is likely that the incremental validity of additional predictors beyond GMA and, say, a structured interview, would be limited.

In fact, even within the context of the present study, some of the estimates of the correlation between supplemental predictors and GMA measures could not be made as precise as would be ideal, at least in comparison to those estimates that are based on the results of major meta-analyses. For example, the job tryout procedure is similar to an extended job sample test. In the absence of data estimating the job tryout–GMA correlation, this correlation was estimated as being the same as the job sample–GMA correlation. It is to be hoped that future research will

provide more precise estimates of this and other correlations between GMA and other personnel measures.

Summary and Implications

Employers must make hiring decisions; they have no choice about that. But they can choose which methods to use in making those decisions. The research evidence summarized in this article shows that different methods and combinations of methods have very different validities for predicting future job performance. Some, such as person-job fit, person-organization fit, and amount of education, have low validity. Others, such as graphology, have essentially no validity; they are equivalent to hiring randomly. Still others, such as GMA tests and integrity tests, have high validity. Of the combinations of predictors examined, two stand out as being both practical to use for most hiring and as having high composite validity: the combination of a GMA test and an integrity test (composite validity of .78); and the combination of a GMA test and a structured interview (composite validity of .76). Both of these combinations can be used with applicants with no previous experience on the job (entry level applicants), as well as with experienced applicants. Both combinations predict performance in job training programs quite well (.78 and .72, respectively), as well as performance on the job. And both combinations are less expensive to use than many other combinations. Hence, both are excellent choices. However, in particular cases there might be reasons why an employer might choose to use one of the other combinations with high, but slightly lower, validity, for example, the combination of a conscientiousness test with a GMA test.

Researchers have used cumulative research findings on the validity of predictors of job performance to create and test theories of job performance. These theories are now shedding light on the psychological processes that underlie observed predictive validity and are advancing

basic understanding of human competence in the workplace.

The validity of the personnel measure (or combination of measures) used in hiring is directly proportional to the practical value of the method—whether measured in dollar value of increased output or percentage increase in output. In economic terms, the gains from increasing the validity of hiring methods can amount over time to literally millions of dollars. However, this can be viewed from the opposite point of view: By using selection methods with low validity, an organization can lose millions of dollars in reduced production, reducing revenue and profits.

In fact, many employers, both in the United States and throughout the world, are currently using suboptimal selection methods. For example, many organizations in France, Israel, and some other countries hire new employees based on handwriting analyses by graphologists. And in the U.S. many organizations rely on measures of "emotional intelligence", person-job fit, or person-organization fit measures. In a competitive world, these organizations are unnecessarily creating a competitive disadvantage for themselves (Schmidt, 1993). By adopting more valid hiring procedures, they could turn this competitive disadvantage into a competitive advantage.

Footnote

1. Recently another procedure for correcting for indirect range restriction has appeared (Le, Oh, Schmidt, & Wooldridge, *in press*). Use of this method in job selection research requires knowledge of an essentially unknowable value, the unrestricted standard deviation (*SD*) of the job performance measure (i.e., the *SD* of job performance in the applicant pool), making it impossible to use in selection research. However, it can be used in many areas of organizational research where estimates of the unrestricted SD of the dependent variable (e.g., job satisfaction or organizational commitment) are available. Simulation studies show it is very accurate in such areas (Le et al., *in press*).

References

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta- analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment, 18*, 291-304.

Arthur, W. Jr., Bell, S. T., Villado, A. J., & Doverspike, D. (2006). The use of person-organization fit in employment decision making: An assessment of its criterion-related validity. *Journal of Applied Psychology, 91,* 786-801.

Arthur, W. Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-153.

Arthur Jr, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435-442.

Bangerter, A., König, C. J., Blatti, S., & Salvisberg, A. (2009). How widespread is graphology in personnel selection practice? A case study of job market myth. *International Journal of Selection and Assessment, 17,* 219–230.

Bar-Hillel, M., & Ben-Shakhar, G. (1986). The a priori case against graphology: Methodological and conceptual issues. In B.Nevo (Ed.), *Scientific aspects of graphology* (pp. 263–279). Springfield, IL: Charles C Thomas.

Ben-Shakhar, G. (1989). Nonconventional methods in personnel selection. In P.Herriot (Ed.), *Handbook of assessment in organizations: Methods and practice for recruitment and appraisal* (pp. 469–485). Chichester, England: Wiley.

Ben-Shakhar, G., Bar-Hillel, M., & Flug, A. (1986). A validation study of graphological

evaluations in personnel selection. In B.Nevo (Ed.), *Scientific aspects of graphology* (pp. 175–191). Springfield, IL: Charles C Thomas.

Ben-Shakhar, G., Bar-Hillel, M., Bilu, Y., Ben-Abba, E., & Flug, A. (1986). Can graphology predict occupational success? Two empirical studies and some methodological ruminations. *Journal of Applied Psychology*, *71*, 645–653.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models evaluating the effects of ratee ability, knowledge, proficiency, temperament, awards, and problem behavior on supervisory ratings. *Journal of Applied Psychology*, *76*, 863–872.

Boudreau, J. W. (1983a). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology*, *36*, 551–576.

Boudreau, J. W. (1983b). Effects of employee flows on utility analysis of human resource productivity improvement programs. *Journal of Applied Psychology*, *68*, 396–406.

Boudreau, J. W. (1984). Decision theory contributions to HRM research and practice. *Industrial Relations*, *23*, 198–217.

Brody, N. (1992). *Intelligence*. New York: Academic Press.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, *2*, 171–183.

Brown, K. G., Le, H., & Schmidt, F. L. (2006). Specific aptitude theory revisited: Is there incremental validity for training performance? *International Journal of Selection and Assessment, 14,* 87 – 100.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Cascio, W. F., & Silbey, V. (1979). Utility of the assessment center as a selection device. *Journal of Applied Psychology*, *64*, 107–118.

Collins, J. M., Schmidt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment, 11*, 17-29.

Collins, J. M., & Schmidt, F. L. (1997). Can suppressor variables enhance criterion-related validity in the personality domain? *Educational and Psychological Measurement, 57,* 924-936.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565-579.

Cooper, M. D. (2001). Job reference immunity statutes: Prevalent but irrelevant. *Cornell Journal of Law and Public Policy, 11,* 1 – 68.

Cronshaw, S. F., & Alexander, R. A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. *Organizational Behavior and Human Decision Processes*, *35*, 102–118.

Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment*, *1*, 153–157.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*, 493–511.

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.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 75*, 639-683.

Hollander, E. P. (1956). The friendship factor in peer nominations. *Personnel Psychology*, *9*, 435–447.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21,* 264 – 276.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, *81*, 459–473.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Department of Labor, Employment Service.

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

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E. A. Fleishman & M. D.Dunnette (Eds.), *Human performance and productivity. Volume I: Human capability assessment* (pp. 233–284). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work force productivity. *American Psychologist*, *38*, 473–478.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in*

research findings. Beverly Hills, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law*, *2*, 447–472.

Hunter, J. E., Schmidt, F. L., & Coggin, T. D. (1988). Problems and pitfalls in using capital budgeting and financial accounting techniques in assessing the utility of personnel programs. *Journal of Applied Psychology*, *73*, 522–528.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721 – 735.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

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

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology*, *75*, 28–42.

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

Jansen, A. (1973). *Validation of graphological judgments: An experimental study*. The Hague, the Netherlands: Monton.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Judge, T. A., Jackson, C. L., Shaw, J. C., Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology, 92*, 107-127.

Joseph, D. L., Jinh, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100*, 298-342.

Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology, 95*, 54-78.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology, 58*, 281-342.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods, 11,* 416 – 438.

Le, H., Oh, I.-S, Schmidt, F. L., & Wooldridge, C. D. (*in press*). Correction for range restriction in meta-analysis revisited: Improvements and implications for organizational research. *Personnel Psychology*.

Levy, L. (1979). Handwriting and hiring. *Dun's Review*, *113*, 72–79.

Marcus, B., Lee, K., & Ashton, M. C. (2007). Personality dimensions explaining relationships between integrity tests and counterproductive behavior: Big five, or one in addition?. *Personnel Psychology, 60*, 1-34.

Matthews, G., Zeidner, M., & Roberts, R. D. (2002). *Emotional intelligence: Science or myth?* Cambridge: MIT Press.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3,* 97 – 105.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology*,

41, 283–314.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Mauer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*, 599–616.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L. & Grubb. W. L., III (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91.

Murphy, K. R. (2006). *A critique of emotional intelligence: What are the problems and how can they be fixed?* London: Lawrence Erlbaum Associates

Neter, E., & Ben-Shakhar, G. (1989). The predictive validity of graphological inferences: A meta-analytic approach. *Personality and Individual Differences*, *10*, 737–745.

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7*, 384-403.

Oh, I.-S., Postlethwaite, B. E., & Schmidt, F. L. (2013). Rethinking the validity of interviews for employment decision making: Implications of recent developments in meta-analysis. In D. J. Svyantek & K. Mahoney (Eds.), *Received wisdom, kernels of truth, and boundary conditions in organizational studies.* Charlotte, NC: Information Age Publishing. Chapter 12, pp. 297 – 329.

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the Five-Factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*, 762-773.

Ones, D. S. (1993). *The construct validity of integrity tests.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*, 679–703.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2012). Integrity tests predict counterproductive work behaviors and job performance well: Comment on Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012). *Journal of Applied Psychology, 97,* 537 – 542.

Passler, K., Beinicke, A., & Hell, B. (2015). Interests and intelligence: A meta-analysis. *Intelligence, 50*, 30-51.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.

Pyburn, K. M., Jr., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*, 143-151.

Rafaeli, A., & Klimoski, R. J. (1983). Predicting sales success through handwriting analysis: An evaluation of the effects of training and handwriting sample content. *Journal of Applied Psychology*, *68*, 212–217.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science*, *1*, 86–89.

Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.

Roth, P. L., BeVier, C. A., Switzer, F. S., & Schippmann, J. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81*(5), 548-556.

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

Rothstein, H. R., & McDaniel, M. A. (1992). Differential validity by sex in employment settings. *Journal of Business and Psychology, 7*, 45-62.

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

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.

Rynes, S. L., Giluk, T. L., & Brown, K. G. (2007). The very separate worlds of academic and practitioner periodicals in human resource management: Implications for evidence-based management. *Academy of Management Journal, 50,* 987 – 1008.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215-227.

Sackett, P. R., & Schmitt, N. (2012). On reconciling conflicting meta-analytic findings regarding integrity test validity. *Journal of Applied Psychology, 97*, 550-556.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302-318.

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*, 299-324.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior*, *33*, 272–292.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schmidt, F. L. (1993). Personnel psychology at the cutting edge. In N. Schmitt & W. Borman (Eds.), *Personnel selection* (pp. 497–515). San Francisco: Jossey Bass.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15,* 187 – 210.

Schmidt, F. L. (2011). A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science, 6,* 560 – 573.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology, 58,* 5 – 9.

Schmidt, F. L., & Hoffman, B. (1973). Empirical comparison of three methods of assessing utility of a selection device. *Journal of Industrial and Organizational Psychology, 1,* 13 – 22.

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. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, *36*, 1128–1137.

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology*, *68*, 407–414.

Schmidt, F. L., & Hunter, J. E. (1992). Development of a causal model of processes determining job performance. *Current Directions in Psychological Science*, *1*, 89–92.

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., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

Schmidt, F. L., & Rothstein, H. R. (1994). Application of validity generalization methods of meta-analysis to biographical data scores in employment selection. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *The biodata handbook: Theory, research, and applications* (pp. 237–260). Palo Alto, CA: Consulting Psychologists Press.

Schmidt, F. L., & Zimmerman, R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89,* 553-561.

Schmidt, F. L., Caplan, J. R., Bemis, S. E., Decuir, R., Dunn, L., & Antone, L. (1979). *Development and evaluation of behavioral consistency method of unassembled*

examining (Tech. Rep. No. 79-21). U.S. Civil Service Commission, Personnel Research and Development Center.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*, 432–439.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, *66*, 166–185.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on workforce productivity. *Personnel Psychology*, *35*, 333–347.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). The impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, *64*, 609–626.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. *Personnel Psychology*, *39*, 1–29.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Goff, S. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology*, *73*, 46–57.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, *32*, 257–281.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology*, *78*, 3–13.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. Park Ranger for three modes of test use. *Journal of Applied Psychology*, *69*, 490–497.

Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59,* 281 – 305.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology*, *43*, 627–670.

Schmidt, F. L., Ones, D. S., & Viswesvaran, C. (1994, June 30–July 3). *The personality characteristic of integrity predicts job training success*. Presented at the 6th Annual Convention of the American Psychological Society, Washington, DC.

Schmidt, F. L., & Rader, M. (1999). Exploring the boundary conditions for interview validity: Meta-analytic validity findings for a new interview type. *Personnel Psychology, 52*, 445-464.

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy of range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61*, 827-868.

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*, 445-494.

Steiner, D. D. (1997). International forum. *The Industrial-Organizational Psychologist*, *34*, 51–53.

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology*, *81*, 134–141.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management*, 29, 609-640.

Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology, 96*, 1167-1194.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012a). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology, 97*, 499-530.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., Odle-Drusseau, H. N. (2012b). The critical role of the research question, inclusion criteria, and transparency in meta-analyses of integrity test research: A reply to Harris et al. (2012) and Ones, Viswesvaran, and Schmidt (2012). *Journal of Applied Psychology, 97*, 543-549.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108-131.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Waters, L. K., & Waters, C. W. (1970). Peer nominations as predictors of short-term sales performance. *Journal of Applied Psychology*, *54*, 42–44.

Wigdor, A. K.Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies*. (Report of the National Research Council Committee on Ability Testing).

Washington, DC: National Academy of Sciences Press.

Table 1

Operational Validity for Overall Job Performance of General Mental Ability (GMA) Combined with a Second Supplementary Predictor Using Multiple Regression

Selection procedures/predictors	Operational validity (r)	Multiple R	Gain in validity (ΔR)	% gain in validity	Standardized regression weights	
					GMA	Supplement
1. GMA tests [a]	.65					
2. Integrity tests [b]	.46	.78	.130	20%	.63	.43
3. Employment interviews (structured) [c]	.58	.76	.117	18%	.52	.43
4. Employment interviews (unstructured) [d]	.58	.73	.087	13%	.49	.38
5. Interests [e]	.31	.71	.062	10%	.64	.29
6. Phone-based interviews (structured) [f]	.46	.70	.057	9%	.56	.29
7. Conscientiousness [g]	.22	.70	.053	8%	.67	.27
8. Reference checks [h]	.26	.70	.050	8%	.65	.26
9. Openness to Experience [i]	.04	.69	.039	6%	.74	-.25
10. Biographical data [j]	.35	.68	.036	6%	.90	-.34
11. Job experience (years) [k]	.16	.68	.032	5%	.66	.21
12. Personality-based EI [l]	.32	.68	.029	5%	.61	.20
13. Person-organization fit [m]	.13	.67	.024	4%	.66	.18
14. SJT (knowledge) [n]	.26	.66	.015	2%	.75	-.17
15. Person-job fit [o]	.18	.66	.014	2%	.64	.13
16. Assessment centers [p]	.36	.66	.013	2%	.78	-.18
17. T & E point method [q]	.11	.66	.009	1%	.65	.11
18. Grade point average [r]	.34	.66	.009	1%	.74	-.14
19. Years of education [s]	.10	.65	.008	1%	.65	.10
20. Extraversion [t]	.09	.65	.006	1%	.65	.09
21. Peer ratings [u]	.49	.65	.006	1%	.57	.12
22. Ability-based EI [v]	.23	.65	.004	0%	.68	-.08
23. Agreeableness [w]	.08	.65	.002	0%	.64	.05
24. Work sample tests [x]	.33	.65	.002	0%	.68	-.06
25. SJT (behavioral tendency) [y]	.26	.65	.001	0%	.64	.03
26. Emotional Stability [z]	.12	.65	.000	0%	.64	.02
27. Graphology [aa]	.02	.65	.000	0%	.65	.02
28. Job tryout procedure [ab]	.44	.65	.000	0%	.63	.02
29. Behavioral consistency method [ac]	.45	.65	.000	0%	.64	.02
30. Job knowledge tests [ad]	.48	.65	.000	0%	.65	-.01
31. Age [ae]	.00	.65	.000	0%	.65	.01

Note. EI = emotional intelligence; SJT = situational judgment tests; T & E = training and experience; In the Table notes, k = the number of studies a meta-analysis is based on.

Selection procedures are listed in the order of gain in operational validity (incremental validity). All values in the third column of Table 1 are operational validities for overall job performance. Unless otherwise noted, all operational validity estimates are corrected for measurement error in the criterion measure and indirect range restriction (IRR) on the predictor measure to estimate operational validity for applicant populations. Details on these operational validities are reported in the Table footnotes. The correlations between GMA and supplementary predictors (used to compute multiple *R*s, gain in validity, and standardized regression weights) are corrected for IRR on GMA but not for measurement error in either measure; these correlations indicate unrestricted observed correlations between the two predictors in applicant populations. Details on these correlations are reported in the Table footnotes.

a From Schmidt, Shaffer, and Oh (2008, Table 3). This operational validity is based on eight individual meta-analytic estimates reported in Table 1 on p. 838. The average across these eight meta-analytic estimates (.65) is presented in Table 3 on p. 843.

b From Ones, Viswesvaran, and Schmidt (1993, Table 8). This operational validity is based on predictive studies conducted on job applicants. The authors of a more recent meta-analysis (Van Iddekinge, Roth, Raymark, and Odle-Dusseau were unable to access many of the studies contained in Ones et al. (1993). We report results based on Ones et al. (1993) because their paper is based on a much larger and more complete set of studies (k = 222 vs. k = 134). The interested readers can refer to a special section published about this research in the *Journal of Applied Psychology* (Ones, Viswesvaran, & Schmidt 2012; Sackett & Schmitt, 2012; Van Idddekinge, Roth, Raymark, & Odel-Dusseau, 2012b). The unrestricted observed correlation between integrity tests and GMA is estimated at .05 (Ones, 1993, Table 30).

c From McDaniel, Whetzel, Schmidt, and Maurer (1994, Table 4, p. 606). This is the operational validity of job-related structured employment interviews based on primary studies in which overall job performance was measured using research-purpose measures (k = 36) and thus represents the most unbiased estimate available. The operational validity presented here was corrected for IRR with the most appropriate meta-analytic reliability estimate for the interview measure from Conway et al. (1995). The unrestricted observed correlation between GMA and the structured interview is .31 (Salgado & Moscoso, 2002, Table 4).

d From McDaniel, Whetzel, Schmidt, and Maurer (1994, Table 4, p. 606). This is the operational validity of job-related unstructured employment interviews based on primary studies in which overall job performance was measured using research-purpose measures (k = 9) and thus represents the most unbiased estimate available. The operational validity presented here was corrected for IRR using the most appropriate meta-analytic reliability estimate for the interview measure from Huffcutt et al. (2013). The unrestricted observed correlation between the unstructured interview and GMA is .41 (Salgado & Moscoso, 2002, Table 3).

e From Van Iddekinge, Roth, Putka, and Lanivich (2011, Table 5, p. 1178). This is the operational validity of job and vocation-focused interest scales that match the dominant interest classification of the job in question. This value was corrected for criterion unreliability and indirect range restriction (k = 26). Because the Nye et al. (2012) meta-analysis was based on fewer studies we did not use it. That meta-analysis used "congruence scores", which are similar to the scale-job matching used in Van Iddekinge et al. Nye et al. obtained a somewhat lower operational validity of .25 averaged across five interest inventories (see their Table 1 results for task performance, p. 390). Nye et al. (2012) corrected for both criterion unreliability and indirect range restriction. The overall, unrestricted correlation between interests and GMA is .04 which is the average correlation between GMA and the six RIASEC interest types (r with Realistic = .20 [ΔR = .026]; r with Investigative = .25 [ΔR = .018]; r with Artistic = -.02 [ΔR = .077]; r with Social = -.16 [ΔR = .125]; r with Enterprising = -.07 [ΔR = .092]; r with Conventional = .01 [ΔR =

.068]). These correlations are from Passler et al. (2015, Table 2). The GMA-interest correlations and corresponding validity increments (ΔR) presented in this note can be used by readers who want to focus on specific RIASEC interest scales. Note that attainment of these incremental validities requires that interest scale and job be matched, as described above.

f From Schmidt and Rader (1999, Table 3, p. 457). This is the operational validity of a non-traditional type of structured employment interview in which questions and correct answers are determined empirically based on their correlation with employee job performance. It is administered via telephone and later scored based on a taped transcript. This type of employment interviews is cost effective because it is conducted via telephone, not face-to-face and "applicants in widely scattered geographical locations can be interviewed with no travel costs for interviewees or interviewers" (Schmidt & Rader, 1999, pp. 450-451). This interview was developed by the Gallup Organization and has been adopted by others. The operational validity reported in this table is for supervisor ratings of job performance (k = 33). Given that it is a type of structured employment interview, the unrestricted observed correlation between GMA and this interview is expected to be .31 (Salgado & Moscoso, 2002, Table 4), as with other structured interviews (refer to note c).

g From Schmidt, Shaffer, and Oh (2008, Table 1 and Appendices C and D). Individual meta-analytic estimates are reported in Table 1 on pp. 838-839 and in Appendix C on pp. 866-867. The averages of these operational validity values (ranging from .04 [Openness] to .22 [Conscientiousness]) are presented in Appendix D on p. 868. We used the average for each personality measure in the current analyses. The gain in validity for Openness (.039) is a bit higher than its operational validity (.036), which is due to its higher correlation with GMA (producing statistical suppression) as compared to the other FFM personality traits. The unrestricted observed correlations between measures of GMA and the FFM personality trait measures are: -.069 (Conscientiousness), .000 (Extraversion), .046 (Agreeableness), Emotional Stability (.159), and Openness to Experience (.380; Judge, Jackson, Shaw, Scott, & Bruce, 2007, Table 3). True score correlations alone were reported in Judge et al. (2007). We attenuated the true score correlations for predictor unreliability in both variables using the psychometric information provided by Timothy A. Judge. It is noted that the operational validity of the FFM traits has been found to be somewhat higher than the values presented here when work-specific, contextualized measures are used (i.e., when items or instructions are specific to work settings) than when general, non-contextualized measured are used (Shaffer & Postlethwaite, 2012, Table 1). The operational validity of the FFM trait personality traits has also been found to be higher when personality traits are measured using ratings by others than by self-reports (Oh, Wang, & Mount, 2011, Table 3). The incremental validities presented in this table are for general (standard) self-report measures of the FFM personality traits and are not these higher values.

h From Hunter and Hunter (1984, Table 9, k = 10). In the absence of any available data, the unrestricted correlation with GMA was assumed to be zero. Assumption of a larger correlation would produce a lower incremental validity.

i Refer to note g.

j From Rothstein, Schmidt, Erwin, Owens, and Sparks (1990, Table 6). The unrestricted observed correlation with GMA is .76 (Schmidt, 1988, p. 283). The standardized regression weight is negative due to its high correlation with GMA (producing statistical suppression). Attainment of the incremental validity presented here is contingent on use of the negative weight on the biodata scale.

k From Hunter and Hunter (1984, Table 6) and Sturman (2003, Table 1). Hunter & Hunter (1984) reported the operational validity of .18 (k = 425) and Sturman (2003, Table 1) reported the operational validity of .13 (k = 68). Thus, the operational validity of job experience (years) is estimated at .16 by

combining these two meta-analytic findings. The unrestricted observed correlation between GMA scores and years of job experience is -.07 (Judge, Jackson, Shaw, Scott, & Bruce, 2007, Table 3).

l From Joseph, Jin, Newman, and O'Boyle (2015, Table 2, p. 307; k = 15). The operational validity of personality traits-based/self-reported EI is corrected for unreliability in the criterion measure (.58, based on Viswesvaran, Schmidt, & Ones, 2005) and indirect range restriction on the predictor measure. The unrestricted correlation between GMA and this type of EI scale is .20 (Joseph & Newman, 2010, Table 2).

m From Arthur, Bell, Villado, and Doverspike (2006, Table 1, k = 36). No correction of this validity for range restriction was possible. Given lack of a meta-analytic estimate for the relationship between P-O Fit and GMA, we performed a meta-analysis in order to derive the estimate to be used in this study. The unrestricted observed correlation between GMA and P-O Fit is -.07 (current study; k = 5); detailed results are available from the authors upon request.

n From McDaniel, Hartman, Whetzel, & Grubb III (2007, Table 3). The operational validities of SJT (knowledge; choosing "the best" or "should do" option; 14 in Table 1) and SJT (behavioral tendency; choosing the "most likely to do" or "would do" option; 25 in Table 1) are the same at .26 (k = 96 and k = 22, respectively). Information needed to correct these validities for any range restriction was not available. The difference in response instructions between these two types of SJTs is discussed in detail in Table 2 of McDaniel et al. (2007). The unrestricted observed correlations between GMA scores and SJT scales are .59 and .36 for SJT (knowledge) and SJT (behavioral tendency), respectively (McDaniel et al., 2007, Table 3). The standardized regression weight for SJT (knowledge) is negative due to its high correlation with GMA, producing statistical suppression); attainment of the small incremental validity presented here is contingent on use of the negative regression weight.

o From Kristof-Brown, Zimmerman, and Johnson (2005, Table 1, k = 19). No correction of this validity for range restriction was possible. Given lack of a meta-analytic estimate for the relationship between P-J fit and GMA, we performed a meta-analysis in order to derive the estimate to be used in this study. The unrestricted observed correlation between GMA and Person-Job Fit is .07 (current study; k = 3); detailed results are available from the authors upon request.

p From Arthur, Day, McNelly, and Edens (2003, Table 3). This is the operational validity averaged for different dimension scores used in assessment centers (i.e., communication, consideration/awareness of others, drive, influencing others, organizing and planning, and problem solving), against the aggregated criterion including supervisor ratings of job performance, promotion, and salary (k = 258). The operational validity of overall assessment center scores for supervisory ratings of job performance is also .36 in Gaugler et al. (1987; Table 8; k = 29). The correlation between GMA and overall assessment center scores is .71 (Collins et al., 2003). The standardized regression weight for the assessment center is negative due to its high correlation with GMA, producing statistical suppression. The incremental validity presented here is contingent on use of this negative weight.

q From McDaniel, Schmidt, & Hunter (1988, Table 3, k = 91). The unrestricted observed correlation between GMA and the T &E point method is taken as zero (Schmidt & Hunter, 1998); this value is based a judgment about the characteristics of this method (i.e., its purely credentialistic nature).

r From Roth, BeVier, Switzer, and Schippmann (1996, Table 2, p. 550). The operational validity estimates for GPA (combination of college, graduate, and PhD/MD GPAs) and college GPA are the same. Here, we report the operational validity of college GPA for job performance given its wider use in employment selection decisions (k = 49). The unrestricted observed correlation between GMA and GPA is .65 (Robbins, Lauver, Le, Davis, Langley, & Carlstrom, 2004, Table 5).

s From Hunter and Hunter (1984, Table 9). For purposes of these calculations, we assumed a zero unrestricted observed correlation between GMA scores and years of education. The reader should remember that this is the correlation within the applicant pool of individuals who apply to get a particular job. In the general population the correlation between education and GMA scores is about .55. Even within applicant pools there is probably at least a small positive correlation. Thus the tiny increment in validity over GMA shown here is probably somewhat of an overestimate.

t Refer to note g.

u From Hunter and Hunter (1984, Tables 8 and 10; $k = 31$). The information needed to correct this validity for range restriction was not available. The average correlation between GMA and peer rating of job performance is approximately .50, which is increased to .65 after correcting for IRR on the GMA measure. If peer ratings are based on an averaged rating from 10 peers, the familiar Spearman-Brown formula indicates that the inter-rater reliability of peer ratings is approximately .90 (Viswesvaran, Ones, & Schmidt, 1996).

v From Joseph, Jin, Newman, and O'Boyle (2015, Table 2, p. 307; $k = 13$). The operational validity of ability-based EI is corrected for unreliability in the criterion measure (assumed to be .58 based on Viswesvaran, Schmidt, & Ones, 2005) and indirect range restriction on the predictor measure. The unrestricted correlation with GMA is .45 (Joseph & Newman, 2010, Table 2). The standardized regression weight is negative due to its high correlation with GMA, producing statistical suppression. Attainment of the incremental validity presented here is contingent on use of this negative regression weight.

w Refer to note g.

x From Roth, Bobko, and McFarland (2005, Table 1, p. 1020; $k = 54$). This is the operational validity for work sample tests for supervisory ratings of job performance. The unrestricted observed correlation of work sample tests with GMA is .57 (Roth et al., 2005, Table 4). The standardized regression weight for work sample tests is negative, albeit small, due to its high correlation with GMA measures, producing statistical suppression; however, work sample tests produce no incremental validity over GMA measures.

y Refer to note n.

z Refer to note g.

aa From Neter and Ben-Shakher (1989), Ben-Shakhar (1989), Ben-Shakhar, Bar-Hillel, Bilu, Ben-Abba, and Flug (1986), and Bar-Hillel and Ben-Shakhar (1986). No correction for range restriction was made on this validity. Range restriction is unlikely here. The unrestricted observed correlation between graphology scores and GMA is assumed to be zero.

ab From Hunter and Hunter (1984, Table 9). No correction of this validity for range restriction could be made. Range restriction is unlikely for the job tryout method, given the absence of any initial screening of applicants. The correlation between job tryout evaluations and GMA scores is estimated at .38 (Schmidt, Hunter, & Outerbridge, 1986); that is, it was taken to be the same as the correlation between job sample tests and GMA. Use of the mean correlation between supervisory performance ratings and GMA scores yields a similar value (.35, uncorrected for measurement error). This correlation increases to .66 after correcting for IRR on the GMA measure.

ac From McDaniel, Schmidt, and Hunter (1988, Table 3; $k = 15$). No information was available that would allow correction of this validity for range restriction. The unrestricted observed correlation between GMA and the behavioral consistency method is .68. This is the expected value given that the achievements measured by this procedure depend substantially on GMA, as well as on personality and other non-cognitive characteristics.

ad From Hunter and Hunter (1984, Table 10; $k = 10$). This validity value could not be corrected for range restriction. The observed correlation between job knowledge scores and GMA scores is .48 (Schmidt, Hunter, & Outerbridge, 1986). The unrestricted observed correlation between job knowledge scores and GMA scores is estimated at .75. Due to its high correlation with GMA scores, the standardized regression weight for job knowledge) is negative, albeit only -.01, producing statistical suppression. Therefore, when using the regression equation presented here, the weight on Job Knowledge would be nearly zero. However, job knowledge measures produce no incremental validity over GMA measures.

ae From Hunter and Hunter (1984, Table 9) and Struman (2003, Table 1). Hunter & Hunter (1984) reported the operational validity of -.01 ($k = 425$) and Sturman (2003, Table 1) reported a very similar finding of .01 ($k = 78$). Thus, the operational validity of age is estimated at .00 by combining these two meta-analytic findings. It is unlikely that this validity is affected by range restriction and no correction was made. The unrestricted observed correlation of age with GMA is assumed to be zero (Schmidt & Hunter, 1998).

Table 2

Operational Validity for Training Performance of General Mental Ability (GMA) Combined with a Second Supplementary Predictor Using Multiple Regression

Selection procedures/predictors	Operational validity (r)	Multiple R	Gain in validity (ΔR)	% gain in validity	Standardized Regression weights	
					GMA	Supple-ment
1. GMA tests [a]	.67					
2. Integrity tests [b]	.43	.78	.109	16%	.65	.40
3. Biographical data [c]	.30	.74	.073	11%	1.04	-.50
4. Employment interviews (unstructured) [d]	.56	.74	.070	11%	.53	.35
5. Interests [e]	.34	.74	.070	11%	.66	.31
6. Conscientiousness [f]	.25	.73	.061	9%	.69	.29
7. Reference checks [g]	.23	.71	.038	6%	.67	.23
8. Employment interviews (structured) [h]	.41	.70	.034	5%	.60	.23
9. Years of education [i]	.20	.70	.029	4%	.67	.20
10. Extraversion [j]	.17	.69	.021	3%	.67	.17
11. Assessment centers [k]	.37	.68	.014	2%	.81	-.20
12. Peer ratings [l]	.36	.68	.008	1%	.76	-.13
13. Agreeableness [m]	.13	.67	.007	1%	.66	.10
14. Emotional Stability [n]	.14	.67	.001	0%	.66	.03
15. Openness to Experience [o]	.24	.67	.000	0%	.67	-.02
16. Job experience (years) [p]	.00	.67	.000	0%	.67	.01

Note. Selection procedures are listed in the order of gain in operational validity. All values in the third column of Table 2 are operational validities for training performance. Unless otherwise noted, all operational validity estimates are corrected for measurement error in the criterion measure and indirect range restriction (IRR) on the predictor measure to estimate operational validity for applicant populations. Details on these operational validities are reported in the Table footnotes. The correlations between GMA and supplementary predictors (used to compute multiple *R*s, gain in validity, and standardized regression weights) are corrected for IRR on GMA but not for measurement error in either measure; these correlations indicate unrestricted observed correlations between the two predictors in applicant populations. Details on these correlations are reported in the Table 1 footnotes.

a From Schmidt, Shaffer, and Oh (2008, Table 3). Individual meta-analytic estimates are reported in Table 2 on p. 840. The average of these estimates across eight meta-analytic estimates (.67) is presented in Table 3 on p. 843. We used this average in the current analyses.

b From Schmidt, Ones, and Viswesvaran (1994). The operational validity reported in this table has been corrected for unreliability in the criterion measure and IRR on the predictor measure. Integrity tests have been found to correlate with GMA scores at .05.

c From Hunter and Hunter (1984, Table 8). The unrestricted observed correlation between biographical data measures and GMA scores is .76 (Schmidt, 1988). The standardized regression weight is negative

due to its high correlation with GMA, producing statistical suppression; attainment of the incremental validity presented here is contingent on use of this negative regression weight.

d From McDaniel, Whetzel, Schmidt, and Maurer (1994, Table 5, p. 606). This is the operational validity of job-related structured employment interviews for training performance $k = 30$). Also refer to note d of Table 1.

e From Van Iddekinge et al. (2011, Table 5, p. 1178). This is the operational validity of job and vocation-focused interest scales that match the dominant interest classification of the job. This was corrected for criterion unreliability and indirect range restriction ($k = 7$). Nye et al.'s (2012) meta-analysis did not examine the criterion of training performance.

f From Schmidt, Shaffer, and Oh (2008; Table 2 and Appendices C and D). Individual meta-analytic validity figures for the Big Five personality measures are reported in their Table 2 on pp. 840-841 and in Appendix C on pp. 866-867. The averages of these estimates are presented in Appendix D on p. 868. We used these averages in the current analyses. See also note g of Table 1.

g From Hunter and Hunter (1984, Table 8; $k = 1$, $N = 1{,}553$). This validity could not be corrected for range restriction. The correlation between reference checks and GMA scores was taken as zero. Assumption of a larger correlation will reduce the estimate of incremental validity.

h From McDaniel, Whetzel, Schmidt, and Maurer (1994, Table 5, p. 606). This is the operational validity of job-related structured employment interviews for training performance ($k = 26$). See also note c of Table 1.

i From Hunter and Hunter (1984). Information was not available that would allow this validity to be corrected for range restriction. For purposes of these calculations, we assumed a zero unrestricted observed correlation between GMA and years of education. The reader should remember that this is the correlation within the applicant pool of individuals who apply to get a particular job. In the general population the correlation between education and GMA scores is about .55. Even within applicant pools there is probably at least a small positive correlation. Thus the increment in validity over GMA shown here is probably somewhat of an overestimate.

j Refer to note f.

k From Gaugler et al. (1987, Table 8). This value of .37 is the operational validity of overall assessment center ratings performance in training ($k = 8$) when corrected for artifacts in the same manner as in Arthur et al. (2003). The standardized regression weight is negative due to its high correlation with GMA, producing statistical suppression. Attainment of the incremental validity shown here requires use of this negative regression weight.

l From Hunter and Hunter (1984, Table 8; $k = 7$). The information needed to correct this validity for range restriction was not available. The average correlation between GMA and peer rating of job performance is approximately .50, which is increased to .65 after correcting for IRR on the GMA measure. If peer ratings are based on an averaged rating from 10 peers, the familiar Spearman-Brown formula indicates that the inter-rater reliability of peer ratings is approximately .90 (Viswesvaran, Ones, & Schmidt, 1996). The standardized regression weight on peer ratings is negative due to its high correlation with GMA, producing statistical suppression. Attainment of the .01 incremental validity presented here requires use of this negative regression weight.

m Refer to note f.

n Refer to note f.

o Refer to note f. The standardized regression weight is negative, albeit very small, due to its high correlation with GMA, producing statistical suppression. However, standard measures of Openness to Experience produce no incremental validity over GMA measures for the prediction of performance in training programs.

p From Hunter and Hunter (1984, Table 6; $k = 90$). These calculations are based the assumption of a zero correlation between years of job experience and GMA measures. The unrestricted observed correlation with GMA is -.07 (Judge, Jackson, Shaw, Scott, & Bruce, 2007).