Psychological Bulletin
1998, Vol. 124, No. 2, 262–274

0033-2909/98/$3.00

The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings

Frank L. Schmidt
University of Iowa

John E. Hunter
Michigan State University

This article summarizes the practical and theoretical implications of 85 years of research in personnel selection. On the basis of meta-analytic findings, this article presents the validity of 19 selection procedures for predicting job performance and training performance and the validity of paired combinations of general mental ability (GMA) and the 18 other selection procedures. Overall, the 3 combinations with the highest multivariate validity and utility for job performance were GMA plus a work sample test (mean validity of .63), GMA plus an integrity test (mean validity of .65), and GMA plus a structured interview (mean validity of .63). A further advantage of the latter 2 combinations is that they can be used for both entry level selection and selection of experienced employees. 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.

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 of learning 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 determined with the aid of 85 years of research. 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; Hunter & Hunter, 1984; Ree & Earles, 1992). 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, measures of conscientiousness and personal integrity, structured employment interviews, and (for experienced workers) job knowledge and work sample tests.

On the basis of meta-analytic findings, this article examines and summarizes what 85 years of research in personnel psychology has revealed about the validity of measures of 19 different selection methods that can be used in making decisions about hiring, training, and developmental assignments. In this sense, this article is an expansion and updating of Hunter and Hunter (1984). In addition, this article examines how well certain combinations of these methods work. These 19 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, for example, and graphology 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. 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 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.

Frank L. Schmidt, Department of Management and Organization, University of Iowa; John E. Hunter, Department of Psychology, Michigan State University.

An earlier version of this article was presented to Korean Human Resource Managers in Seoul, South Korea, June 11, 1996. The presentation was sponsored by Tong Yang Company. We would like to thank President Wang-Ha Cho of Tong Yang for his support and efforts in this connection. We would also like to thank Deniz Ones and Kuh Yoon for their assistance in preparing Tables 1 and 2 and Gershon Ben-Shakhar for his comments on research on graphology.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management and Organization, College of Business, University of Iowa, Iowa City, Iowa 52240. Electronic mail may be sent to frank-schmidt@uiowa.edu.

Research over the last 15 years 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 $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 large 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 actual formula for computing practical gains per person hired per year on the job is a three way product (Brogden, 1949; Schmidt et al., 1979):

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

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

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

(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 method and the old selection method. 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}. $\bar{Z}_x$ 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 (1982a, 1982b), 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 methods, 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: 85 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, 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; Hunter, Schmidt, & Jackson, 1982). 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 fact, 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 (Hunter, 1980; Schmidt, Hunter, & Pearlman, 1980). 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 article.

Table 1 summarizes research findings for the prediction of performance on the job. The first column of numbers in Table 1 shows the estimated mean validity of 19 selection methods for predicting performance on the job, as revealed by meta-analyses conducted over the last 20 years. 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.

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 first column in Table 2 shows the mean validity of 10 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.

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 to range restriction on the selection method in incumbent samples relative to applicant populations. Observed validity estimates so corrected estimate operational validities of selection methods when used to hire from applicant pools. Operational validities are also referred to as true validities.

In the pantheon of 19 personnel measures in Table 1, GMA (also called general cognitive ability and general intelligence) occupies a special place, for several reasons. First, of all procedures that can be used for all jobs, whether entry level or advanced, it has the highest validity and lowest application cost. Work sample measures are slightly more valid but are much more costly and can be used only with applicants who already know the job or have been trained for the occupation or job. Structured employment interviews are more costly and, in some forms, contain job knowledge components and therefore are not suitable for inexperienced, entry level applicants. The assessment center and job tryout are both much more expensive and have less validity. 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 & Hunter, 1981). Literally thousands of studies have been conducted over the last nine decades. By contrast, only 89 validity studies of the structured interview have been conducted (McDaniel, Whetzel, Schmidt, & Mauer, 1994). 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 in job training programs (Hunter, 1986; Hunter & Hunter, 1984; Ree & Earles, 1992). 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 over 90 years (Brody, 1992; Carroll, 1993; Jensen, 1998). As a result of this massive related research literature, the meaning of the construct of intelligence is much clearer than, for example, the meaning of what is measured by interviews or assessment centers (Brody, 1992; Hunter, 1986; Jensen, 1998).

The value of .51 in Table 1 for the validity of GMA is from a very large meta-analytic study conducted for the U.S. Department of Labor (Hunter, 1980; Hunter & Hunter, 1984). The database for this unique meta-analysis included over 32,000 employees in 515 widely diverse civilian jobs. This meta-analysis examined both performance on the job and performance in job training programs. This meta-analysis found that the validity of GMA for predicting job performance was .58 for professional-managerial jobs, .56 for high level complex technical jobs, .51 for medium complexity jobs, .40 for semi-skilled jobs, and .23 for completely unskilled jobs. The validity for the middle complexity level of jobs (.51)—which includes 62% of all

Table 1
Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores Combined With a Second Predictor Using (Standardized) Multiple Regression

Personnel measures	Validity (r)	Multiple R	Gain in validity from adding supplement	% increase in validity	Standardized regression weights	
					GMA	Supplement
GMA tests[a]	.51					
Work sample tests[b]	.54	.63	.12	24%	.36	.41
Integrity tests[c]	.41	.65	.14	27%	.51	.41
Conscientiousness tests[d]	.31	.60	.09	18%	.51	.31
Employment interviews (structured)[e]	.51	.63	.12	24%	.39	.39
Employment interviews (unstructured)[f]	.38	.55	.04	8%	.43	.22
Job knowledge tests[g]	.48	.58	.07	14%	.36	.31
Job tryout procedure[h]	.44	.58	.07	14%	.40	.20
Peer ratings[i]	.49	.58	.07	14%	.35	.31
T & E behavioral consistency method[j]	.45	.58	.07	14%	.39	.31
Reference checks[k]	.26	.57	.06	12%	.51	.26
Job experience (years)[l]	.18	.54	.03	6%	.51	.18
Biographical data measures[m]	.35	.52	.01	2%	.45	.13
Assessment centers[n]	.37	.53	.02	4%	.43	.15
T & E point method[o]	.11	.52	.01	2%	.39	.29
Years of education[p]	.10	.52	.01	2%	.51	.10
Interests[q]	.10	.52	.01	2%	.51	.10
Graphology[r]	.02	.51	.00	0%	.51	.02
Age[s]	−.01	.51	.00	0%	.51	−.01

Note. T & E = training and experience. The percentage of increase in validity is also the percentage of increase in utility (practical value). All of the validities presented are based on the most current meta-analytic results for the various predictors. See Schmidt, Ones, and Hunter (1992) for an overview. All of the validities in this table are for the criterion of overall job performance. Unless otherwise noted, all validity estimates are corrected for the downward bias due to measurement error in the measure of job performance and range restriction on the predictor in incumbent samples relative to applicant populations. The correlations between GMA and other predictors are corrected for range restriction but not for measurement error in either measure (thus they are smaller than fully corrected mean values in the literature). These correlations represent observed score correlations between selection methods in applicant populations.

[a] From Hunter (1980). The value used for the validity of GMA is the average validity of GMA for medium complexity jobs (covering more than 60% of all jobs in the United States). Validities are higher for more complex jobs and lower for less complex jobs, as described in the text. [b] From Hunter and Hunter (1984, Table 10). The correction for range restriction was not possible in these data. The correlation between work sample scores and ability scores is .38 (Schmidt, Hunter, & Outerbridge, 1986). [c,d] From Ones, Viswesvaran, and Schmidt (1993, Table 8). The figure of .41 is from predictive validity studies conducted on job applicants. The validity of .31 for conscientiousness measures is from Mount and Barrick (1995, Table 2). The correlation between integrity and ability is zero, as is the correlation between conscientiousness and ability (Ones, 1993; Ones et al., 1993). [e,f] From McDaniel, Whetzel, Schmidt, and Mauer (1994, Table 4). Values used are those from studies in which the job performance ratings were for research purposes only (not administrative ratings). The correlations between interview scores and ability scores are from Huffcutt, Roth, and McDaniel (1996, Table 3). The correlation for structured interviews is .30 and for unstructured interviews, .38. [g] From Hunter and Hunter (1984, Table 11). The correction for range restriction was not possible in these data. The correlation between job knowledge scores and GMA scores is .48 (Schmidt, Hunter, & Outerbridge, 1986). [h] From Hunter and Hunter (1984, Table 9). No correction for range restriction (if any) could be made. (Range restriction is unlikely with this selection method.) The correlation between job tryout ratings and ability scores is estimated at .38 (Schmidt, Hunter, & Outerbridge, 1986); that is, it was taken to be the same as that between job sample tests and ability. Use of the mean correlation between supervisory performance ratings and ability scores yields a similar value (.35, uncorrected for measurement error). [i] From Hunter and Hunter (1984, Table 10). No correction for range restriction (if any) could be made. The average fully corrected correlation between ability and peer ratings of job performance is approximately .55. If peer ratings are based on an average rating from 10 peers, the familiar Spearman–Brown formula indicates that the interrater reliability of peer ratings is approximately .91 (Viswesvaran, Ones, & Schmidt, 1996). Assuming a reliability of .90 for the ability measure, the correlation between ability scores and peer ratings is $.55\sqrt{.91(.90)} = .50$. [j] From McDaniel, Schmidt, and Hunter (1988a). These calculations are based on an estimate of the correlation between T & E behavioral consistency and ability of .40. This estimate reflects the fact that the achievements measured by this procedure depend on not only personality and other noncognitive characteristics, but also on mental ability. [k] From Hunter and Hunter (1984, Table 9). No correction for range restriction (if any) was possible. In the absence of any data, the correlation between reference checks and ability was taken as .00. Assuming a larger correlation would lead to lower estimated incremental validity. [l] From Hunter (1980), McDaniel, Schmidt, and Hunter (1988b), and Hunter and Hunter (1984). In the only relevant meta-analysis, Schmidt, Hunter, and Outerbridge (1986, Table 5) found the correlation between job experience and ability to be .00. This value was used here. [m] The correlation between biodata scores and ability scores is .50 (Schmidt, 1988). Both the validity of .35 used here and the intercorrelation of .50 are based on the Supervisory Profile Record Biodata Scale (Rothstein, Schmidt, Erwin, Owens, and Sparks, 1990). (The validity for the Managerial Profile Record Biodata Scale in predicting managerial promotion and advancement is higher [.52; Carlson, Scullen, Schmidt, Rothstein, & Erwin, 1998]. However, rate of promotion is a measure different from overall performance on one's current job and managers are less representative of the general working population than are first line supervisors). [n] From Gaugler, Rosenthal, Thornton, and Benson (1987, Table 8). The correlation between assessment center ratings and ability is estimated at .50 (Collins, 1998). It should be noted that most assessment centers use ability tests as part of the evaluation process; Gaugler et al. (1987) found that 74% of the 106 assessment centers they examined used a written test of intelligence (see their Table 4). [o] From McDaniel, Schmidt, and Hunter (1988a, Table 3). The calculations here are based on a zero correlation between the T & E point method and ability; the assumption of a positive correlation would at most lower the estimate of incremental validity from .01 to .00. [p] From Hunter and Hunter (1984, Table 9). For purposes of these calculations, we assumed a zero correlation between years of education and ability. 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 ability is about .55. Even within applicant pools there is probably at least a small positive correlation; thus, our figure of .01 probably overestimates the incremental validity of years of education over general mental ability. Assuming even a small positive value for the correlation between education and ability would drive the validity increment of .01 toward .00. [q] From Hunter and Hunter (1984, Table 9). The general finding is that interests and ability are uncorrelated (Holland, 1986), and that was assumed to be the case here. [r] From Neter and Ben-Shakhar (1989), Ben-Shakhar (1989), Ben-Shakhar, Bar-Hillel, Bilu, Ben-Abba, and Flug (1986), and Bar-Hillel and Ben-Shakhar (1986). Graphology scores were assumed to be uncorrelated with mental ability. [s] From Hunter and Hunter (1984, Table 9). Age was assumed to be unrelated to ability within applicant pools.

Table 2
Predictive Validity for Overall Performance in Job Training Programs of General Mental Ability (GMA) Scores Combined With a Second Predictor Using (Standardized) Multiple Regression

Personnel measures	Validity (r)	Multiple R	Gain in validity from adding supplement	% increase in validity	Standardized regression weights	
					GMA	Supplement
GMA Tests[a]	.56					
Integrity tests[b]	.38	.67	.11	20%	.56	.38
Conscientiousness tests[c]	.30	.65	.09	16%	.56	.30
Employment interviews (structured and unstructured)[d]	.35	.59	.03	5%	.59	.19
Peer ratings[e]	.36	.57	.01	1.4%	.51	.11
Reference checks[f]	.23	.61	.05	9%	.56	.23
Job experience (years)[g]	.01	.56	.00	0%	.56	.01
Biographical data measures[h]	.30	.56	.00	0%	.55	.03
Years of education[i]	.20	.60	.04	7%	.56	.20
Interests[j]	.18	.59	.03	5%	.56	.18

Note. The percentage of increase in validity is also the percentage of increase in utility (practical value). All of the validities presented are based on the most current meta-analytic results reported for the various predictors. All of the validities in this table are for the criterion of overall performance in job training programs. Unless otherwise noted, all validity estimates are corrected for the downward bias due to measurement error in the measure of job performance and range restriction on the predictor in incumbent samples relative to applicant populations. All correlations between GMA and other predictors are corrected for range restriction but not for measurement error. These correlations represent observed score correlations between selection methods in applicant populations.

[a] The validity of GMA is from Hunter and Hunter (1984, Table 2). It can also be found in Hunter (1980). [b,c] The validity of .38 for integrity tests is from Schmidt, Ones, and Viswesvaran (1994). Integrity tests and conscientiousness tests have been found to correlate zero with GMA (Ones, 1993; Ones, Viswesvaran & Schmidt, 1993). The validity of .30 for conscientiousness measures is from the meta-analysis presented by Mount and Barrick (1995, Table 2). [d] The validity of interviews is from McDaniel, Whetzel, Schmidt, and Mauer (1994, Table 5). McDaniel et al. reported values of .34 and .36 for structured and unstructured interviews, respectively. However, this small difference of .02 appears to be a result of second order sampling error (Hunter & Schmidt, 1990, Ch. 9). We therefore used the average value of .35 as the validity estimate for structured and unstructured interviews. The correlation between interviews and ability scores (.32) is the overall figure from Huffcutt, Roth, and McDaniel (1996, Table 3) across all levels of interview structure. [e] The validity for peer ratings is from Hunter and Hunter (1984, Table 8). These calculations are based on an estimate of the correlation between ability and peer ratings of .50. (See note i to Table 1). No correction for range restriction (if any) was possible in the data. [f] The validity of reference checks is from Hunter and Hunter (1984, Table 8). The correlation between reference checks and ability was taken as .00. Assumption of a larger correlation will reduce the estimate of incremental validity. No correction for range restriction was possible. [g] The validity of job experience is from Hunter and Hunter (1984, Table 6). These calculations are based on an estimate of the correlation between job experience and ability of zero. (See note l to Table 1). [h] The validity of biographical data measures is from Hunter and Hunter (1984, Table 8). This validity estimate is not adjusted for range restriction (if any). The correlation between biographical data measures and ability is estimated at .50 (Schmidt, 1988). [i] The validity of education is from Hunter and Hunter (1984, Table 6). The correlation between education and ability within applicant pools was taken as zero. (See note p to Table 1). [j] The validity of interests is from Hunter and Hunter (1984, Table 8). The correlation between interests and ability was taken as zero (Holland, 1986).

the jobs in the U.S. economy—is the value entered in Table 1. This category includes skilled blue collar jobs and mid-level white collar jobs, such as upper level clerical and lower level administrative jobs. Hence, the conclusions in this article apply mainly to the middle 62% of jobs in the U.S. economy in terms of complexity. The validity of .51 is representative of findings for GMA measures in other meta-analyses (e.g., Pearlman et al., 1980) and it is a value that produces high practical utility.

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. In the U.S. Department of Labor research, the average predictive validity performance in job training programs was .56 (Hunter & Hunter, 1984, Table 2); this is the figure entered in Table 2. 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. (As can be seen from Table 2, this is also true of integrity tests, conscientiousness tests, and employment interviews.)

Because of its special status, GMA can be considered the primary personnel measure for hiring decisions, and one can consider the remaining 18 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 .51 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 of increase in validity produced by the adding the second measure is also the percentage of increase in practical value (utility). 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 correlations 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 appropriately represent the observed score correlations between predictors in applicant populations, we corrected all correlations between GMA and other predictors for range restriction but not for measurement error in the measure of either predictor.

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. Work sample tests are often used to hire skilled workers, such as welders, machinists, and carpenters. When combined in a standardized regression equation with GMA, the work sample receives a weight of .41 and GMA receives a weight of .36. (The standardized regression weights are given in the last two columns of Tables 1 and 2.) The validity of this weighted sum of the two measures (the multiple R) is .63, which represents an increment of .12 over the validity of GMA alone. This is a 24% increase in validity over that of GMA alone—and also a 24% increase in the practical value (utility) of the selection procedure. As we saw earlier, this can be expressed as a 24% increase in the gain in dollar value of output. Alternatively, it can be expressed as a 24% increase in the percentage of increase in output produced by using GMA alone. In either case, it is a substantial improvement.

Work sample tests can be used only with applicants who already know the job. 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.

Integrity tests are used in industry to hire employees with reduced probability of counterproductive job behaviors, such as drinking or drugs on the job, fighting on the job, stealing from the employer, sabotaging equipment, and other undesirable behaviors. They do predict these behaviors, but they also predict evaluations of overall job performance (Ones, Viswesvaran, & Schmidt, 1993). Even though their validity is lower, integrity tests produce a larger increment in validity (.14) and a larger percentage of increase in validity (and utility) than do work samples. This is because integrity tests correlate zero with GMA (vs. .38 for work samples). In terms of basic personality traits, integrity tests have been found to measure mostly conscientiousness, but also some components of agreeableness and emotional stability (Ones, 1993). The figures for conscientiousness measures per se are given in Table 1. The validity of conscientiousness measures (Mount & Barrick, 1995) is lower than that for integrity tests (.31 vs. .41), its increment to validity is smaller (.09), and its percentage of increase in validity is smaller (18%). However, these values for conscientiousness are still large enough to be practically useful.

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 .38 (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 20% 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. However, the increment in validity produced by measures of conscientiousness (.09, for a 16% increase) is only slightly smaller. The validity estimate for conscientiousness is based on 21 studies and 4,106 individuals (Mount & Barrick, 1995), a somewhat larger database.

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. As a result, structured interviews are more costly to construct and use, but are also more valid. As shown in Table 1, the average validity of the structured interview is .51, versus .38 for the unstructured interview (and undoubtedly lower for carelessly conducted unstructured interviews). An equally weighted combination of the structured interview and a GMA measure yields a validity of .63. As is the case for work sample tests, the increment in validity is .12 and the percentage of increase is 24%. These figures are considerably smaller for the unstructured interview (see Table 1). Clearly, the combination of a structured interview and a GMA test is an attractive hiring procedure. It achieves 63% of the maximum possible practical value (utility), and does so at reasonable cost.

As shown in Table 2, both structured and unstructured interviews predict performance in job training programs with a validity of about .35 (McDaniel et al., 1994; see their Table 5). The incremental validity for the prediction of training performance is .03, a 5% increase.

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 increase the validity by .07 over that of GMA measures alone, yielding a 14% increase in validity and utility. Thus job knowledge tests can have substantial practical value to the organization using them.

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, little validity information is available for this criterion, and there is no entry in Table 2 for job knowledge tests.

The next three personnel measures in Table 1 increase validity and utility by the same amount as job knowledge tests (i.e., 14%). However, two of these methods are considerably less practical to use in many situations. Consider the job tryout procedure. Unlike job knowledge 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 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 (and incremental validity), 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 destroying 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 be terminated after a period of time anyway.

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. Like the job tryout procedure, peer ratings have some limitations. First, 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).

The behavioral consistency method of evaluating previous training and experience (McDaniel, Schmidt, & Hunter, 1988a; 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, 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.

In terms of time and cost, the behavioral consistency procedure is nearly as time consuming and costly to construct as locally constructed job knowledge tests. 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 job knowledge tests and other 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.

No information is available on the validity of the job tryout or the behavioral consistency procedures for predicting performance in training programs. However, as indicated in Table 2, peer ratings have been found to predict performance in training programs with a mean validity of .36 (see Hunter & Hunter, 1984, Table 8).

For the next procedure, reference checks, 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. 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. That is, past employers today 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.

However, in recent years the legal climate in the United States has been changing. Over the last decade, 19 of the 50 states have enacted laws that provide immunity from legal liability for employers providing job references in good faith to other employers, and such laws are under consideration in 9 other

states (Baker, 1996). Hence, reference checks, formerly a heavily relied on procedure in hiring, may again come to provide an increment to the validity of a GMA measure for predicting job performance. In Table 1, the increment is 12%, only two percentage points less than the increments for the five preceding methods.

Older research indicates that reference checks predict performance in training with a mean validity of .23 (Hunter & Hunter, 1984, Table 8), yielding a 9% increment in validity over GMA tests, as shown in Table 2. But, again, these findings may no longer hold; however, changes in the legal climate may make these validity estimates accurate again.

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 .18 and the increment in validity (and utility) over that from GMA alone is only .03 (a 6% increase). However, Schmidt, Hunter, and Outerbridge (1986) found that when 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 a job will predict job performance only for the first 5 years on the job. 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. Hunter and Hunter (1984, Table 6) reported a mean validity of .01. 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.

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. 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, 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 but produce an increment in validity over GMA of only .01 on average (a 2% increase). The reason that the increment in validity is so small is that biographical data correlates substantially with GMA (.50; 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 (Hunter & Hunter, 1984, Table 8). However, because of their relatively high correlation with GMA, they produce no increment in validity for performance in training.

Biographical data measures are technically difficult and time consuming to construct (although they are easy to use once constructed). Considerable statistical sophistication is 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.

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 and business games. 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.

Assessment centers are like biographical data measures: They have substantial validity but only moderate incremental validity over GMA (.01, a 2% increase). The reason is also the same: They correlate moderately highly with GMA—in part because they typically include a measure of GMA (Gaugler et al., 1987). 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.

Assessment centers have generally not been used to predict performance in job training programs; hence, their validity for this purpose is unknown. However, assessment center scores do predict rate of promotion and advancement in management. Gaugler et al. (1987, Table 8) reported a mean validity of .36 for this criterion (the same value as for the prediction of job

performance). Measurements of career advancement include number of promotions, increases in salary over given time spans, absolute level of salary attained, and management rank attained. Rapid advancement in organizations requires rapid learning of job related knowledge. Hence, assessment center scores do appear to predict the acquisition of job related knowledge on the job.

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., 1988a), 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 relevant schooling (or each course taken), 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 are measured. As shown in Table 1, the T&E point method has low validity and produces only a 2% increase in validity over that available from GMA alone. The T&E point method has not been used to predict performance in training programs.

Sheer amount of education has even lower validity for predicting job performance than the T&E point method (.10). However, its increment to validity, rounded to two decimal places, is the same .01 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 semi-skilled 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 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.

Many believe that interests are an important determinant of one's level of job performance. People whose interests match the content of their jobs (e.g., people with mechanical interests who have mechanical jobs) are believed to have higher job performance than with nonmatching interests. The validity of .10 for interests shows that this is true only to a very limited extent. To many people, this seems counterintuitive. Why do interests predict job performance so poorly? Research indicates that interests do substantially influence which jobs people prefer and which jobs they attempt to enter. However, once individuals are in a job, the quality and level of their job performance is determined mostly by their mental ability and by certain personality traits such as conscientiousness, not by their interests. So despite popular belief, measurement of work interests is not a good means of predicting who will show the best future job performance (Holland, 1986).

Interests predict learning in job training programs somewhat better than they predict job performance. As shown in Table 2, Hunter and Hunter (1984, Table 8) found a mean validity of .18 for predicting performance in job training programs.

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) and in Israel. 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 nongraphologists can infer some (limited) information about their personalities and job performance from the handwriting samples. But untrained nongraphologists do just as well as graphologists; both show validities in the .18–.20 range. When the assessees are required to copy the same material from a book to create their handwriting sample, there is no evidence that graphologists or nongraphologists can infer any valid information about personality traits or job performance from the handwriting samples (Neter & Ben-Shakhar, 1989). What this indicates 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, expressions of emotions, verbal fluency, grammatical skills, and so on. 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 another counterintuitive finding, like the finding that interests are a poor predictor of job performance. 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 the 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.

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 again that job experience, which is positively correlated with age, is not correlated with performance in training programs (see Table 2).

Finally, we address an issue raised by a reviewer. 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 firms. The reviewer therefore questioned whether it was 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 the standard deviations of validity across studies (applications), after the appropriate corrections for sampling error and other statistical and measurement artifacts, are 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 (see the next section), but whatever those combinations are, they do not appear to vary much from one application (study) to another. 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 in different organizations and settings.

Toward a Theory of the Determinants of Job Performance

The previous section 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 eight decades and involving millions of employees. 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. In recent years, the focus in personnel psychology has turned to the 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 change of emphasis 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 article. 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 interview scores and scores on GMA tests is .32 (Huffcutt et al., 1996). 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 mental ability 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, mental ability 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 nonsupervisory jobs, this direct effect is only about 20% as large as the indirect effect; 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 may be 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 (.65) for predicting job performance. Another combination with high validity (.63) 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).

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, and it would be informative to examine the incremental validity from adding a third predictor. 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 intercorrelations in the literature makes this impossible at the present time. In the future, as data accumulates, such analyses may become feasible.

In fact, even within the context of the present study, some of the estimated predictor intercorrelations 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–ability test score correlation, this correlation was estimated as being the same as the job sample–ability test 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.

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 are obviously relevant here.

For differential validity, the general finding has been that validities (the focus of this study) do not differ appreciably for different subgroups. For predictive fairness, the usual finding has been a lack of predictive bias for minorities and women. That is, given similar scores on selection procedures, later job performance is similar regardless of group membership. 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 (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). That is, the predictive interpretation of scores is the same in different subgroups.

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 interests and amount of education, have very low validity. Others, such as graphology, have essentially no validity; they are equivalent to hiring randomly. Still others, such as GMA tests and work sample measures, 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 .65); and the combination of a GMA test and a structured interview (composite validity of .63). 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 (.67 and .59, 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. Some examples are combinations that include

conscientiousness tests, work sample tests, job knowledge tests, and the behavioral consistency method.

In recent years, 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 of 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.

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 other countries hire new employees based on handwriting analyses by graphologists. And many organizations in the United States rely solely on unstructured interviews, when they could use more valid methods. 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.

References

Baker, T. G. (1996). Practice network. *The Industrial-Organizational Psychologist, 34,* 44–53.

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., 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.

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.

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 or utility analysis of human resources productivity improvement programs. *Journal of Applied Psychology, 68,* 396–407.

Boudreau, J. W. (1984). Decision theory contributions to human resource management 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.

Carlson, K. D., Scullen, S. E., Schmidt, F. L., Rothstein, H. R., & Erwin, F. W. (1998). *Generalizable biographical data: Is multi-organizational development and keying necessary?* Manuscript in preparation.

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. (1998). *Prediction of overall assessment center evaluations from ability, personality, and motivation measures: A meta-analysis.* Unpublished manuscript, Texas A & M University, College Station, TX.

Cronshaw, S. F., & Alexander, R. A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. *Organizational Behavior and Human Performance, 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., & Benson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Holland, J. (1986). New directions for interest testing. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 245–267). Hillsdale, NJ: Erlbaum.

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

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. (1982a). Fitting people to jobs: Implications of personnel selection for 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. (1982b). 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., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, 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.

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.

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

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988a). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988b). Job experience correlates of job performance. *Journal of Applied Psychology, 73,* 327–330.

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.

Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 13, pp. 153–200). JAI Press.

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

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 Monograph, 78,* 679–703.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology, 65,* 373–407.

Rafaeli, A., & Klimoski, R. J. (1983). Predicting sales success through handwriting analysis: An evaluation of the effects of training and handwriting sample context. *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.

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.

Schmidt, F. L. (1988). The problem of group differences in ability 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., Caplan, J. R., Bemis, S. E., Decuir, R., Dinn, 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. (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,* 407–415.

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

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. (1986). The 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., Outerbridge, A. N., & Goff, S. (1988). The joint relation of experience and ability with job performance: A test of three hypotheses. *Journal of Applied Psychology, 73,* 46–57.

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

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1980). Task difference and validity of aptitude tests 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., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian Validity Generalization Model. *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., 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., & 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.

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.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–560.

Waters, L. K., & Waters, C. W. (1970). Peer nominations as predictors of short-term role 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.

Received April 8, 1997
Revision received February 3, 1998
Accepted April 2, 1998 ■