Psychological Bulletin
1984, Vol. 96, No. 1, 72-98

Validity and Utility of Alternative Predictors of Job Performance

John E. Hunter
Michigan State University

Ronda F. Hunter
College of Education
Michigan State University

Meta-analysis of the cumulative research on various predictors of job performance shows that for entry-level jobs there is no predictor with validity equal to that of ability, which has a mean validity of .53. For selection on the basis of current job performance, the work sample test, with mean validity of .54, is slightly better. For federal entry-level jobs, substitution of an alternative predictor would cost from $3.12 billion (job tryout) to $15.89 billion per year (age). Hiring on ability has a utility of $15.61 billion per year, but affects minority groups adversely. Hiring on ability by quotas would decrease this utility by 5%. A third strategy—using a low cutoff score—would decrease utility by 83%. Using other predictors in conjunction with ability tests might improve validity and reduce adverse impact, but there is as yet no data base for studying this possibility.

A crucial element in the maintenance of high productivity in both government and private industry is the selection of people with high ability for their jobs. For most jobs, the only presently known predictive devices with high validity are cognitive and psychomotor ability tests. Recent work on validity generalization (Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1977; Schmidt & Hunter, 1980; Schmidt, Hunter, Pearlman, & Shane, 1979) has shown that most findings of low validity are due to artifacts, mainly statistical error due to small sample size. Similar and broader conclusions follow from reanalyses of past meta-analyses by Lent, Aurbach, and Levin (1971), as noted by Hunter (1980b) and from inspection of unpublished large-sample studies done in the United States Army by Helm, Gibson, and Brogden (1957), in Hunter (1980c), and in Schmidt, Hunter, and Pearlman (1981). High selection validity translates into considerable financial savings for most organizations. Hunter (1979a) estimated that if the Philadelphia, Pennsylvania, Police Department were to drop its use of a cognitive ability test to select entry-level officers, the cost to the city would be more than $170 million over a 10-year period. Schmidt, Hunter, McKenzie, and Muldrow (1979) show that over a 10-year period, with a 50% selection ratio, the federal government could save $376 million by using a cognitive ability test to select computer programmers rather than by selecting at random.[1]

Hunter and Schmidt (1982b) applied a utility model to the entire national work force. Gains from using cognitive tests rather than selecting at random are not as spectacular at the national level as might be predicted by findings from single organizations. This is because high-ability people who are selected for crucial top-level jobs will not be available for lower-level jobs, where they would bring higher productivity. However, even considering this cancellation of effect, Hunter and Schmidt estimated in 1980 that productivity differences between complete use of cognitive ability tests and no use of the tests would amount to a minimum of $80 billion per year. That is, productivity differences due to use or nonuse of present tests would be, in the current state of the economy, about as great as total corporate profits, or about 20% of the total federal budget.

To replace cognitive ability tests with any instrument of lower validity would incur very high costs, because even minute differences in

Requests for reprints should be sent to John E. Hunter, Department of Psychology, Michigan State University, East Lansing, Michigan 48824.

[1] Random selection, though rarely if ever practiced, provides a base line for comparison of utility gains and losses that best demonstrates the potential magnitude of the differences.

validity translate into large dollar amounts. These costs would be shared by everyone, regardless of sex or group affiliation, because failure to increase productivity enough to compensate for inflation and to compete in foreign markets affects the entire economy.

Adverse Impact and Test Fairness

Unfortunately, the use of cognitive ability tests presents a serious problem for American society; there are differences in the mean ability scores of different racial and ethnic groups that are large enough to affect selection outcomes. In particular, blacks score, on the average, about one standard deviation lower than whites, not only on tests of verbal ability, but on tests of numerical and spatial ability as well (e.g., see U.S. Employment Service, 1970). Because fewer black applicants achieve high scores on cognitive ability tests, they are more likely to fall below selection cutoff scores than are white applicants. For example, if a test is used to select at a level equivalent to the top half of white applicants, it will select only the top 16% of the black applicants. This difference is *adverse impact* as defined by the courts.

Fifteen years ago the elimination of adverse impact seemed a straightforward but arduous task: It was a question of merely amending the tests. Testing professionals reasoned that if there were no mean difference between racial groups in actual ability, the mean difference in test scores implied that tests were unfair to black applicants. If tests were unfair to black applicants, then it was only necessary to remove all test content that was culturally biased. Not only would adverse impact then vanish, but also the validity of the test would increase. Moreover, if a test were culturally unfair to blacks, it would probably prove to be culturally unfair to disadvantaged whites as well, and reforming the tests would eliminate that bias also.

However, the empirical evidence of the last 15 years has not supported this reasoning (Schmidt & Hunter, 1981). Evidence showing that single-group validity (i.e., validity for one group but not for others) is an artifact of small sample sizes (Boehm, 1977; Katzell & Dyer, 1977; O'Connor, Wexley, & Alexander, 1975; Schmidt, Berner, & Hunter, 1973) has shown that any test that is valid for one racial group is valid for the other. Evidence of differential validity (i.e., validity differences between subgroups) to be an artifact of small sample size (Bartlett, Bobko, Mosier, & Hannan, 1978; Hunter, Schmidt, & Hunter, 1979) suggests that validity is actually equal for blacks and whites. Finally, there is an accumulation of evidence directly testing for cultural bias, showing results that are consistently opposite of that predicted by the test bias hypothesis. If test scores for blacks were lower than their true ability scores, then their job performance would be higher than that predicted by their test scores. In fact, however, regression lines for black applicants are either below or equal to the regression lines for white applicants (see review studies cited in Hunter, Schmidt, & Rauschenberger, in press; Schmidt & Hunter, 1980, 1981). This finding holds whether job performance measures are ratings or objective measures of performance, such as production records or job sample measures.

The evidence is clear: The difference in ability test scores is mirrored by a corresponding difference in academic achievement and in performance on the job. Thus, the difference in mean test scores reflects a real difference in mean developed ability. If the difference is the result of poverty and hardship, then it will vanish as poverty and hardship are eliminated. However, because the difference currently represents real differences in ability, construction of better tests will not reduce adverse impact. In fact, better tests, being somewhat more reliable, will have slightly more adverse impact.

Adverse impact can be eliminated from the use of ability tests, but only by sacrificing the principle that hiring should be in order of merit. If we hire solely on the basis of probable job performance, then we must hire on ability from the top down, without regard to racial or ethnic identification. The result will be lower hiring rates for minority applicants. If we decide to moderate adherence to the merit principle so as to apply the principle of racial and ethnic balance, then the method that loses the least productivity in the resulting work force is that of hiring on the basis of ability from the top down within each ethnic group separately. This method produces exact quotas while maximizing the productivity of the workers selected within each ethnic group.

Many organizations are presently using a

third method. They use ability tests, but with a very low cutoff score. This method does reduce adverse impact somewhat, but it reduces the labor savings to the organization by far more (Hunter, 1979a, 1981a; Mack, Schmidt, & Hunter, undated). The low-cutoff method is vastly inferior to quotas in terms of either productivity or ethnic balance. This point will be considered in detail in the section on economic analysis.

Adverse Impact Gap and the Promise of Alternative Predictors

As long as there is a true difference in ability between groups, there will be a gap in their relative performance that can never be completely closed. The fact that tests are fair means that differences in mean test score between groups are matched by differences in mean job performance. However, differences in test score are not matched by equivalent differences in job performance, because the differences are affected by the validity of the test. Figure 1 illustrates this point: If the predicted criterion score (e.g., performance rating) for any test score is the same for all applicants, and if test validity is .50 (the validity for most jobs after correction for restriction in range and unreliability of criterion measures), then a difference of one standard deviation on the mean test score corresponds to half a standard deviation's difference on performance.

If the test score cutoff is set to select the top half of the white applicants, 50% will perform above the mean on the criterion. At the same cutoff, only 16% of the black applicants will be selected, although 31% would perform at or above the mean for white applicants. Improving the test cannot entirely overcome this difference. If the test is improved so that there is half a standard deviation difference on both test and criterion, 50% of whites as against 31% of blacks will be selected. (This is the true meaning of Thorndike's, 1971, claim that a test that is fair to individuals might be unfair to groups.)

Although the gap cannot be entirely closed, it can be narrowed by increasing the validity of the selection process. If we could find predictors of determinants of performance other than ability to add to the prediction supplied by ability tests, then we could simultaneously increase validity and decrease adverse impact. That is, the most likely successful approach to reduced adverse impact is not through the discovery of substitutes for ability tests—there is no known test that approaches cognitive tests in terms of validity for most jobs—but through the discovery of how best to use alternative measures in conjunction with cognitive ability tests.

Distinguishing Method From Content

In considering the use of alternatives, especially their use in conjunction with ability tests, it is helpful to distinguish the means of measurement (method) from the specification of what is to be measured (content). Most contemporary discussion of alternative predictors of job performance is confused by the failure to make this distinction. Much contemporary work is oriented toward using ability to predict performance but using something other than a test to assess ability. The presumption is that an alternative measure of ability might find less adverse impact. The existing cumulative literature on test fairness shows this to be a false hope. The much older and larger literature on the measurement of abilities also suggested that this would be a false hope. All large-sample studies through the years have shown that paper-and-pencil tests are excellent measures of abilities and that other kinds of tests are usually more expensive and less valid. Perhaps an investigation of characteristics other than ability would break new ground, if content were to be considered apart from method.

An example of confusion between content and method is shown in a list of alternative predictors in an otherwise excellent recent review published by the Office of Personnel Management (Personnel Research and Development Center, 1981). This review lists predictors such as reference checks, self-assessments, and interviews as if they were mutually exclusive. The same is true for work samples, assessment centers, and the behavioral consistency approach of Schmidt, Caplan, et al., 1979). However, reference checks often ask questions about ability, as tests do; about social skills, as assessment centers do; or about past performance, as experience ratings do. Self-assessments, interviews, and tests may all be used to measure the same char-

acteristics. A biographical application blank may seek to assess job knowledge by asking about credentials, or it may seek to assess social skill by asking about elective offices held.

What characteristics could be assessed that might be relevant to the prediction of future job performance? The list includes: past performance on related jobs; job knowledge; psychomotor skills; cognitive abilities; social skills; job-related attitudes such as need for achievement (Hannan, 1979), locus of control (Hannan, 1979), or bureaucratic value orientation (Gordon, 1973); emotional traits such as resistance to fear or stress or to enthusiasm. Such characteristics form the content to be measured in predicting performance.

What are existing measurement methods? They come in two categories: methods entailing observation of behavior, and methods entailing expert judgment. Behavior is observed by setting up a situation in which the applicant's response can be observed and measured. The observation is valid to the extent that the count or measurement is related

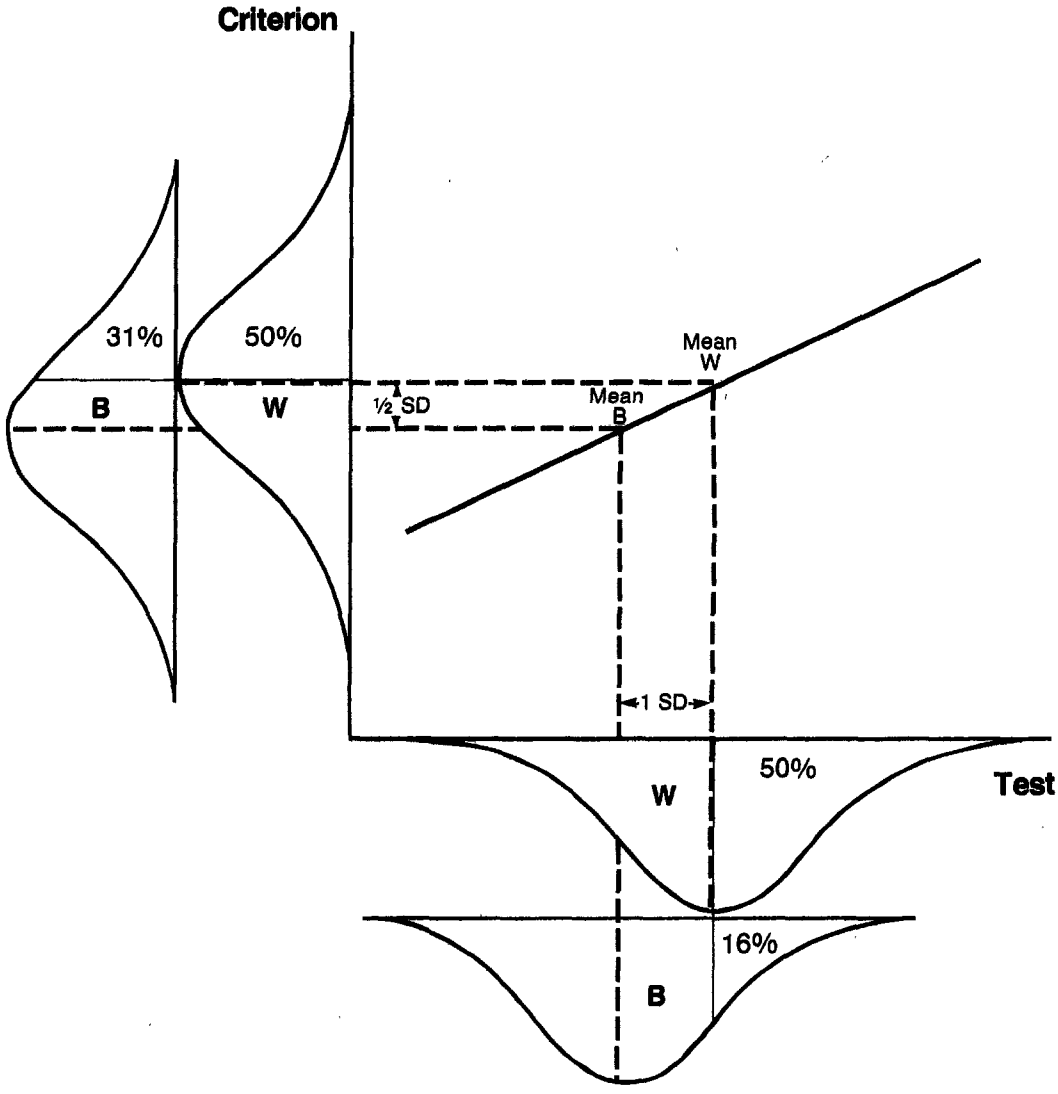

Figure 1. This shows that a fair test may still have an adverse impact on the minority group. (From "Concepts of Culture Fairness" by R. L. Thorndike, 1971, *Journal of Educational Measurement, 8,* pp. 63–70. Copyright 1971 by the National Council on Measurement in Education. Reprinted by permission.)

to the characteristic. Observation of behavior includes most tests, that is, situations in which the person is told to act in such a way so as to produce the most desirable outcome. The response may be assessed for correctness, speed, accuracy, or quality. Behavior may also be observed by asking a person to express a choice or preference, or to make a judgment about some person, act, or event. A person's reports about memories or emotional states can also be considered as observations, although their validity may be more difficult to demonstrate. If judgment is to be used as the measurement mode, then the key question is "Who is to judge?" There are three categories for judges: self-assessments, judgments by knowledgeable others such as supervisor, parent, or peer, and judgments by strangers such as interviewers or assessment center panels. Each has its advantages and disadvantages. The data base for studies of judgment is largest for self-judgments, second largest for judgments by knowledgeable others, and smallest for judgments by strangers. However, the likelihood of distortion and bias in such studies is in the same order.

Figure 2 presents a scheme for the classification of potential predictors of job performance. Any of the measurement procedures along the side of the diagram can be used to assess any of the characteristics along the top. Of course the validity of the measurement may vary from one characteristic to another, and the relevance of the characteristics to particular jobs will vary as well.

Meta-Analysis and Validity Generalization

In the present study of alternative predictors of job performance, the results of hundreds of studies were analyzed using methods called

METHOD (Means of Measurement)			CONTENT (Aspect of Behavior or Thought)						
Category		Method	Performance	Knowledge	Ability	Psychomotor Skill	Social Skill	Attitude	Emotional Trait
Observation	Informal, on the job	Inspection & Appraisal							
		Incidence Counts							
		Coding of Company Records							
	Controlled Measurement (Tests)	Speed							
		Accuracy							
		Quality							
		Adaptiveness & Innovation							
		Choice							
Judgment	Judgment of Traits	Self							
		Informed Other							
		Stranger							
	Judgment of Relevant Experience	Self							
		Informed Other							
		Stranger							

Figure 2. Measurement content and method: A two-dimensional structure in which various procedures are applied to the assessments of various aspects of thought and behavior.

validity generalization in the area of personnel selection and called *meta-analysis* in education or by people using the particular method recommended by Glass (1976). The simplest form of meta-analysis is to average correlations across studies, and this was our primary analytic tool. However, we have also used the more advanced formulas from personnel selection (Hunter, Schmidt, & Jackson, 1982), because they correct the variance across studies for sampling error. Where possible we have corrected for the effects of error of measurement and range restriction. Because validity generalization formulas are not well known outside the area of personnel selection, we review the basis for such corrections in this section.

The typical method of analyzing a large number of studies that bear on the same issue is the narrative review. The narrative review has serious drawbacks, however (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Jackson, 1980). Much better results can be obtained by quantitative analysis of results across studies. Glass (1976) coined the word meta-analysis for such quantitative analysis. A variety of meta-analysis methods have been devised, including the counting of positive and negative results, the counting of significance tests, and the computing of means and variances of results across studies. The strengths and weaknesses of the various methods are reviewed in Hunter, Schmidt, and Jackson (1982).

The most common method of meta-analysis is that which was proposed by Glass (1976), and was presented in detail in Glass, McGaw, and Smith (1981). The steps in this method are:

1. Gather as many studies on the issue as possible.
2. Code the data so that the outcome measure in each study is assessed by the same statistic.
3. Assess the general tendency in the findings by averaging the outcome statistic across studies.
4. Analyze the variance across studies by breaking the studies down into relevant subgroups and comparing the means of the subgroups.

In the area of personnel selection, the Glass (1976) method has been in use for at least 50 years (Dunnette, 1972; Ghiselli, 1966, 1973; Thorndike, 1933). Ghiselli gathered hundreds of validation studies on all methods of predicting job performance. Each validation study was coded to yield a validity coefficient, that is, a Pearson correlation coefficient between the predictor of job performance (such as an ability test score) and the measure of job performance (such as a rating of performance by the worker's immediate supervisor). Ghiselli then averaged validity coefficients across studies (although he used the median instead of the mean). He analyzed the variance of validity across studies by breaking the studies down into families of jobs and comparing average validity across job families.

The Glass (1976) method has one main deficiency. If the variance across studies is taken at face value, then the problem of sampling error is neither recognized nor dealt with adequately (Hedges, 1983; Hunter, 1979b; Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 1977). Schmidt and Hunter (1977) showed that ignoring sampling error leads to disastrous results in the area of personnel selection. Because he ignored the effect of sampling error on the variance of findings across studies, Ghiselli (1966, 1973) concluded that tests are only valid on a sporadic basis, that validity varies from one setting to another because of subtle differences in job requirements that have not yet been discovered. He therefore concluded that no predictor could be used for selection in any given setting without justifying its use with a validation study conducted in that exact setting.

The following example shows how different conclusions may be drawn when sampling error is considered. In the present study we found that the average population correlation for the interview as a predictor of supervisor ratings of performance is .11 (.14 if corrected for error of measurement in ratings). For simplicity, let us suppose that the population correlation for the interview in every setting is .11. Let us also suppose that all studies on the interview were done with a sample size of 68 (Lent, Aurbach, & Levin, 1971, found that the average sample size across more than 1,500 studies was 68). Then the average observed correlation across studies would be close to .11 (depending on the total sample size across studies), but the standard deviation would be

far from zero; the observed standard deviation would be .12. The correlations would be spread over a range from −.25 to +.47. A little over 16% of the correlations would be negative, and a little over 16% would be .20 or greater.

The correct conclusion from this example is, "The validity of the interview is .11 in all settings." However, if the variance of results across studies is taken at face value, as it is in Glass's (1976) method, then the conclusion is, "The validity of the interview varies enormously from one setting to the next. In about one sixth of settings, however, the validity reaches a reasonable level of .20 or better." If the number of studies is small, then some feature of the studies would by chance be correlated with the variation in observed correlations, and the reviewer would say something like, "Studies done with female interviewers were more likely to produce useful levels of validity." If the number of studies is large, then no study feature would be correlated with outcome and the conclusion would probably read, "The validity of the interview varies according to some subtle feature of the organization which is not yet known. Further research is needed to find out what these moderator variables are."

Use of the significance test in this context adds to the confusion. The correlation would be significant in only 14% of the studies (that is, it would be wrong in 86% of the studies, reflecting the usual low level of power is psychological studies in which the null hypothesis is false), leading to the false conclusion, "The interview was found to be valid in only 14% of studies—9% greater than chance—and thus cannot be used except in that 9% of settings in which a local validation study has shown it to be valid."

It is possible to correct the variance across studies to eliminate the effect of sampling error. Schmidt and Hunter (1977) devised a formula for doing so under the rubric of *validity generalization,* though that first formula had an error that was corrected in Schmidt, Hunter, Pearlman, and Shane (1979). Schmidt and Hunter and others (Callender & Osburn, 1980; Hedges, 1983; Raju & Burke, in press) have developed a variety of formulas since then. The historical basis for these formulas is reviewed in Hunter, Schmidt, and Pearlman (1982), and the mathematical basis for comparing them is given in Hunter, Schmidt, and Jackson (1982). In the context of personnel selection, the formulas all give essentially identical results.

Schmidt and Hunter (1977) noted that study outcomes are distorted by a number of artifacts other than sampling error. They presented meta-analysis formulas that correct for error of measurement and range restriction as well as sampling error. The comparable formulas of Hunter, Schmidt, and Jackson (1982) were used when applicable in this study. However, there are artifacts that are not controlled by even these formulas, the most important of which are bad data, computational errors, transcriptional errors, and typographical errors. Thus, even the corrected standard deviations reported in the present study are overestimates of the actual variation in validity across jobs.

Productivity Gains and Losses

A further advantage of meta-analysis is that the validity information from the meta-analysis can be combined with estimated costs to form a utility analysis for the use of alternative predictors in place of written tests. Dollar impact figures can be generated in a form appropriate for use by individual firms or organizations. The cost of testing is a significant part of this impact. Because the cost of paper-and-pencil tests is negligible in comparison with the gains to be obtained from a selection process, this factor has usually been ignored in the literature. Cost is a more significant factor, however, for assessment centers, miniature training and evaluation, and probationary periods; taking cost into consideration adds to the accuracy of utility estimates for all predictors.

Design of This Study

This study was begun with the intent of using meta-analyses to study the cumulative research on alternative predictors in four ways: validity of the alternative predictor considered alone and in conjunction with ability tests, and adverse impact of the predictor considered alone and in conjunction with ability tests. However, we could discover a data base for only one of the four: validity of the predictor

considered alone. Few studies consider more than one predictor on the same data set. Furthermore, most of the studies of multiple predictors do not present the correlations between the predictors, so that multiple regression cannot be done. The lack of data on adverse impact has to do with the age of the studies. Because early studies found low validity for alternative predictors, research on alternative predictors tended to die out by the time that analysis by racial group began to become routine. The sparse literature that does exist is summarized in Reilly and Chao (1982).

In all the new meta-analyses conducted for this study, it was possible to correct the variance of validity for sampling error. Corrections for reliability and range restriction were made when the necessary data were available, as noted below. Correction for error of measurement in job performance was done using the upper bound interrater reliability of .60 for ratings (King, Hunter, & Schmidt, 1980) and a reliability of .80 for training success as measured by tests. Except for ability, correlations are not corrected for restriction in range. Empirical values for range restriction are known only for ability tests, and most of the alternative predictors considered here are relatively uncorrelated with cognitive ability. Also, when Hunter (1980c) analyzed the United States Employment Service data base, he found incumbent-to-applicant standard deviation ratios of .67, .82, and .90 for cognitive, perceptual, and psychomotor ability, respectively. The figure of .90 for psychomotor ability suggests that restriction in range on alternative predictors stems largely from indirect restriction due to selection using ability.

There are two predictors for which restriction in range may be a problem: the interview and ratings of training and experience. For example, some of the studies of the interview might have been done in firms using the interview for selection, in which case restriction in range was present. The key finding for the interview is a mean validity of .14 (in predicting supervisor ratings of performance, corrected for error of measurement). If the restriction in range is as high as that for cognitive ability, then the true mean validity of the interview is .21. The observed average correlation for ratings of training and experience is .13. If the extent of restriction in range matches that for cognitive ability, then the correct correlation would be .19.

No correction for restriction in range was needed for biographical inventories because it was always clear from context whether the study was reporting tryout or follow-up validity, and follow-up correlations were not used in the meta-analyses.

Studies analyzed. This study summarizes results from thousands of validity studies. Twenty-three new meta-analyses were done for this review. These analyses, reported in Table 8, review work on six predictors: peer ratings, biodata, reference checks, college grade point average (GPA), the interview, and the Strong Interest Blank. These meta-analyses were based on a heterogeneous set of 202 studies: 83 predicting supervisor ratings, 53 predicting promotion, 33 predicting training success, and 33 predicting job tenure.

Table 1 reanalyzes the meta-analysis done by Ghiselli (1966, 1973). Ghiselli summarized thousands of studies done over the first 60 years of this century. His data were not available for analysis using state-of-the-art methods.

Tables 1 through 7 review meta-analyses done by others. Table 2 contains a state-of-the-art meta-analysis using ability to predict supervisor ratings and training success. This study analyzed the 515 validation studies carried out by the United States Employment Service using its General Aptitude Test Battery (GATB). Tables 3 through 6 review comparative meta-analyses by Dunnette (1972), Reilly and Chao (1982), and Vineberg and Joyner (1982), who reviewed 883, 103, and 246 studies, respectively. Few of the studies reviewed by Reilly and Chao date as far back as 1972; thus, their review is essentially independent of Dunnette's. Vineberg and Joyner reviewed only military studies, and cited few studies found by Reilly and Chao. Thus, the Vineberg and Joyner meta-analysis is largely independent of the other two. Reilly's studies are included in the new meta-analyses presented in Table 8.

Tables 6 and 7 review meta-analyses of a more limited sort done by others. Table 6 presents unpublished results on age, education, and experience from Hunter's (1980c) analysis of the 515 studies done by the United States Employment Service. Table 6 also presents meta-analyses on job knowledge (Hunter,

1982), work sample tests (Asher & Sciarrino, 1974), and assessment centers (Cohen, Moses, & Byham, 1974), reviewing 16, 60, and 21 studies, respectively.

Validity of Alternative Predictors

This section presents the meta-analyses that cumulate the findings of hundreds of validation studies. First, the findings on ability tests are reviewed. The review relies primarily on a reanalysis of Ghiselli's lifelong work (Ghiselli, 1973), and on the first author's recent meta-analyses of the United States Employment Service data base of 515 validation studies (Hunter, 1980c). Second, past meta-analyses are reviewed, including major comparative reviews by Dunnette (1972), Reilly and Chao (1982), and Vineberg and Joyner (1982). Third, new meta-analyses are presented using the methods of validity generalization. Finally, there is a summary comparison of all the predictors studied, using supervisor ratings as the measure of job performance.

Predictive Power of Ability Tests

There have been thousands of studies assessing the validity of cognitive ability tests. Validity generalization studies have now been run for over 150 test-job combinations (Brown, 1981; Callender & Osburn, 1980, 1981; Lilienthal & Pearlman, 1983; Linn, Harnisch, & Dunbar, 1981; Pearlman, Schmidt, & Hunter, 1980; Schmidt et al., 1980; Schmidt & Hunter, 1977; Schmidt, Hunter, & Caplan, 1981a, 1981b; Schmidt, Hunter, Pearlman, & Shane, 1979; Sharf, 1981). The results are clear: Most of the variance in results across studies is due to sampling error. Most of the residual variance is probably due to errors in typing, computation, or transcription. Thus, for a given test-job combination, there is essentially no variation in validity across settings or time. This means that differences in tasks must be large enough to change the major job title before there can be any possibility of a difference in validity for different testing situations.

The results taken as a whole show far less variability in validity across jobs than had been anticipated (Hunter, 1980b; Pearlman, 1982; Schmidt, Hunter, & Pearlman, 1981). In fact three major studies applied most known methods of job classification to data on training success; they found no method of job analysis that would construct job families for which cognitive ability tests would show a significantly different mean validity. Cognitive ability has a mean validity for training success of about .55 across all known job families (Hunter, 1980c; Pearlman, 1982; Pearlman & Schmidt, 1981). There is no job for which cognitive ability does not predict training success. Hunter (1980c) did find, however, that psychomotor ability has mean validities for training success that vary from .09 to .40 across job families. Thus, validity for psychomotor ability can be very low under some circumstances.

Recent studies have also shown that ability tests are valid across all jobs in predicting job proficiency. Hunter (1981c) recently reanalyzed Ghiselli's (1973) work on mean validity. The major results are presented in Table 1, for Ghiselli's job families. The families have been arranged in order of decreasing cognitive complexity of job requirements. The validity of cognitive ability decreases correspondingly, with a range from .61 to .27. However, even the smallest mean validity is large enough to generate substantial labor savings as a result of selection. Except for the job of sales clerk, the validity of tests of psychomotor ability increases as job complexity decreases. That is, the validity of psychomotor ability tends to be high on just those job families where the validity of cognitive ability is lowest. Thus, multiple correlations derived from an optimal combination of cognitive and psychomotor ability scores tend to be high across all job families. Except for the job of sales clerk, where the multiple correlation is only .28, the multiple correlations for ability range from .43 to .62.

There is a more uniform data base for studying validity across job families than Ghiselli's heterogeneous collection of validation studies. Over the last 35 years, the United States Employment Service has done 515 validation studies all using the same aptitude battery. The GATB has 12 tests and can be scored for 9 specific aptitudes (U.S. Employment Service, 1970) or 3 general abilities. Hunter (1980a, 1980b; 1981b, 1981c) has analyzed this data base and the main results are presented in Table 2. After considering 6 major

Table 1
Hunter's (1981c) Reanalysis of Ghiselli's (1973) Work on Mean Validity

Job families	Mean validity			Beta weight		
	Cog	Per	Mot	Cog	Mot	*R*
Manager	.53	.43	.26	.50	.08	.53
Clerk	.54	.46	.29	.50	.12	.55
Salesperson	.61	.40	.29	.58	.09	.62
Protective professions worker	.42	.37	.26	.37	.13	.43
Service worker	.48	.20	.27	.44	.12	.49
Trades and crafts worker	.46	.43	.34	.39	.20	.50
Elementary industrial worker	.37	.37	.40	.26	.31	.47
Vehicle operator	.28	.31	.44	.14	.39	.46
Sales clerk	.27	.22	.17	.24	.09	.28

Note. Cog = general cognitive ability; Per = general perceptual ability; Mot = general psychomotor ability; *R* = multiple correlation. Mean validities have been corrected for criterion unreliability and for range restriction using mean figures for each predictor from Hunter (1980c) and King, Hunter, and Schmidt (1980).

methods of job analysis, Hunter found that the key dimension in all methods was the dimension of job complexity. This dimension is largely captured by Fine's (1955) data dimension, though Fine's things dimension did define 2 small but specialized industrial job families: set-up work and jobs involving feeding/offbearing. (Fine developed scales for rating all job according to their demands for dealing with data, things, and people.) The present analysis therefore groups job families by level of complexity rather than by similarity of tasks. Five such job families are listed in Table 2, along with the percentage of U.S. workers in each category. The principal finding for the prediction of job proficiency was that although the validity of the GATB tests varies across job families, it never approaches zero. The validity of cognitive ability as a predictor decreases as job complexity decreases. The validity of psychomotor ability as a predictor increases as job complexity decreases. There is corresponding variation in the beta weights for the regression equation for each job family as shown in Table 2. The beta weights for cognitive ability vary from .07 to .58; the beta weights for psychomotor ability vary from .00 to .46. The multiple correlation varies little across job families, from .49 to .59 with an average of .53.

The results of all cumulative analyses are congruent. If general cognitive ability alone is used as a predictor, the average validity across all jobs is .54 for a training success criterion and .45 for a job proficiency criterion. Regarding the variability of validities across jobs, for a normal distribution there is no limit as to how far a value can depart from the mean. The probability of its occurrence just decreases from, for example, 1 in 100 to 1 in 1,000 to 1 in 10,000, and so forth. The phrase *effective range* is used for the interval that captures the most values. In the Bayesian literature, the convention is 10% of cases. For tests of cognitive ability used alone, then, the effective range of validity is .32 to .76 for training success and .29 to .61 for job proficiency. However, if cognitive ability tests are combined with psychomotor ability tests, then the average validity is .53, with little variability across job families. This very high level of validity is the standard against which all alternative predictors must be judged.

Past Comparisons of Predictors

Three meta-analytic studies (Reilly & Chao, 1982; Dunnette, 1972; Vineberg & Joyner, 1982) have compared the validity of different predictors; two studies were completed after the present study began. All three studies have been reanalyzed for summary here.

Table 3 presents a summary of the findings of Dunnette (1972). The predictors are presented in three groups. Group 1 shows Dunnette's data as reanalyzed by the present authors in terms of three general abilities tests rather than in terms of many specific aptitudes.

Table 2
Relationship Between Ability as Tested by the GATB and Job Performance for Jobs at Five Levels of Complexity

Level	Dimension	Complexity family	% U.S. workers	Mean validity for training success			Mean validity for performance			Regression equation to predict performance			Multiple *R*
				GVN	SPQ	KFM	GVN	SPQ	KFM	GVN	SPQ	KFM	
1	Things	Setting up	2.5	.65	.53	.09	.56	.52	.30	.40	.19	.07	.59
2	Data	Synthesizing/coordinating	14.7	.50	.26	.13	.58	.35	.21	.58			.58
3	Data	Analyzing/compiling/computing	62.7	.57	.44	.31	.51	.40	.32	.45		.16	.53
4	Data	Comparing/copying	17.7	.54	.53	.40	.40	.35	.43	.28		.33	.50
5	Things	Feeding/offbearing	2.4				.23	.24	.48	.07		.46	.49

Note. Based on data from the U.S. Employment Service (Hunter, 1980c). Dimensions and complexity families from Fine (1955). GVN = cognitive ability; SPQ = perceptual ability; KFM = psychomotor ability; GATB = General Aptitude Test Battery.

The resulting validities are comparable with those found in the United States Employment Service data base as shown in Table 2. Group 2 includes three predictors that might be used for entry-level jobs. Of these, only biodata has a validity high enough to compare with that of ability. The interview is a poor predictor, and amount of education as a predictor did not work at all for the jobs Dunnette was interested in. Group 3 includes two predictors that can be used only if examinees have been specifically trained for the job in question: job knowledge and job tryout. The validity of .51 for job knowledge is comparable with the validity of .53 for ability tests. The validity of .44 for job tryout is lower than one might expect. The most likely cause of the low validity of job tryout is the halo effect in supervisor ratings. Because of the halo effect, if the performance of the worker on tryout was evaluated by a supervisor rating, then the upper bound on the reliability of the tryout measurement would be .60 (King et al., 1980). In that case, error in supervisor ratings reduces the reliability of the predictor rather than that of the criterion. If so, then job tryout with perfect measurement of job performance might have a validity of .57 (corrected using a reliability of .60) rather than the observed validity of .44.

Table 4 presents a summary of the findings of Reilly and Chao (1982). Most of their findings are included in the meta-analyses presented in Table 8. The validity of .38 for biodata is comparable with the .34 found by Dunnette (1972), and the finding of .23 for the interview is roughly comparable with the .16 found by Dunnette. The finding of .17 for academic achievement, however, is much higher than the zero for education Dunnette found. Dunnette used amount of education as the predictor, however, whereas Reilly and Chao used grades. One would expect grades to be more highly correlated with cognitive ability than amount of education. For that reason, grades would have higher validity than amount of education.

Reilly and Chao's (1982) estimate of some validity for self-assessments is probably an overestimate. Under research conditions people can provide reasonable self-assessments on dimensions—for example typing speed (Ash, 1978, $r = .59$), spelling (Levine, Flory, & Ash,

Table 3
Meta-Analysis Derived From Dunnette (1972)

Predictors	No. correlations	Average validity
Group 1		
Cognitive ability	215	.45
Perceptual ability	97	.34
Psychomotor ability	95	.35
Group 2		
Biographical inventories	115	.34
Interviews	30	.16
Education	15	.00
Group 3		
Job knowledge	296	.51
Job tryout	20	.44

1977, $r = .58$), and intelligence (DeNisi & Shaw, 1977, $r = .35$)—on which they received multiple formal external assessments. However, there is a drastic drop in validity for self-assessments of dimensions on which persons have not received external feedback. Some examples are typing statistical tables (Ash, 1978, $r = .07$), filing (Levine, Flory, & Ash, 1977, $r = .08$), and manual speed and accuracy (DeNisi & Shaw, 1977, $r = .19$). As the assessment shifts from task in school to task on job, the validity drops to very low levels. Furthermore, these correlations are misleadingly high because they are only indirect estimates of validity. For example, if a person's self-assessment of intelligence correlates .35 with his or her actual intelligence, and if intelligence correlates .45 with job performance, then according to path analysis the correlation between self-assessment and job performance would be about $.35 \times .45 = .16$. This correlation would be even lower if self-assessments in hiring contexts proved to be less accurate than self-assessments under research conditions.

We consider four studies that actually used self-assessments in validation research. Bass and Burger (1979) presented international comparisons of the validity of self-assessments of performance in management-oriented situational tests measuring such traits as leadership style and attitude toward subordinates. The average correlation with advancement, across a total sample size of 8,493, was .05, with no variation after sampling error was eliminated. Johnson, Guffey, and Perry (1980) correlated the self-assessments of welfare counselors with supervisor ratings and found a correlation of $-.02$ (sample size = 104). Van Rijn and Payne (1980) correlated self-estimates on 14 dimensions of fire fighting with objective measures such as work sample tests. Only 2 of the 14 correlations were significant, and both were negative (sample size = 209). Farley and Mayfield (1976) correlated self-assessments with later performance ratings for 1,128 insurance salespersons and found no significant relation (hence, $r = .05$ for this sample size). The average validity of self-assessments across these 4 studies is zero. Self-assessments appear to have no validity in operational settings.

Reilly and Chao (1982) also found that projective tests and handwriting analysis are not valid predictors of job performance.

Table 5 presents the summary results of the Vineberg and Joyner (1982) study, corrected for error of measurement in ratings using a reliability of .60. These researchers reviewed only military studies. The most striking feature of Table 5 is the low level of the validity coefficients. The rank order of the correlations for supervisor rating criteria is about the same as in other studies (i.e., training performance, aptitude, biodata, education, interest), but the level is far lower for all variables except education. This fact may reflect a problem with military operational performance ratings, evidence for which can be found in one of the Vineberg and Joyner tables. They found that if a performance test was used as the criterion, then the validity was more than twice as high as when performance ratings were used as the criterion. Many of the ratings were actually

Table 4
Meta-Analysis Derived From Reilly and Chao (1982)

Predictor	No. correlations	Average validity
Biographical inventory	44	.38
Interview	11	.23
Expert recommendation	16	.21
Reference check	7	.17
Academic achievement	10	.17
Self-assessment	7	Some
Projective tests	5	Little
Handwriting analysis	3	None

Table 5
Meta-Analysis of Military Studies Using Supervisory Ratings as the Criterion

Predictor	Performance measures: Global rating	Suitability	All ratings
	Number of correlations		
Aptitude	101	11	112
Training performance	51	7	58
Education	25	10	35
Biographical inventory	12	4	16
Age		10	10
Interest	15		15
	Average validity[a]		
Aptitude	.21	.35	.28
Training performance	.27	.29	.28
Education	.14	.36	.25
Biographical inventory	.20	.29	.24
Age		.21	.21
Interest	.13		.13

Note. Derived from Vineberg and Joyner (1982).
[a] Global ratings, suitability, and all ratings were corrected for error of measurement using a reliability of .60.

suitability ratings (the military term for ratings of potential ability). The difference between suitability ratings and performance ratings is similar to the differences between ratings of potential and ratings of performance that has shown up in assessment center studies, and is discussed later.

Table 6 presents the results for four other meta-analytic studies. The data on age, education, and experience were taken from the United States Employment Service data base (Hunter, 1980c). The validity results for age in Table 6 were gathered on samples of the general working population. Thus, these samples would have included few teenagers. However, there have been military studies that reported noticeably lower performance in recruits aged 17 and younger. That is, although there were no differences as a function of age for those 18 years and older, those 17 years of age and younger were less likely to complete training courses. This was usually attributed to immaturity rather than to lack of ability.

Hunter's (1982) job knowledge study largely involved studies in which job knowledge was used as a criterion measure along with work sample performance or supervisor ratings. The validity for job knowledge tests is high—.48 with supervisor ratings and .78 with work sample performance tests. However, the use of job knowledge tests is limited by the fact that they can only be used for prediction if the examinees are persons already trained for the job.

The same can be said for the work sample tests considered by Asher and Sciarrino (1974). These tests are built around the assumption that the examinee has been trained to the task. In fact, the verbal work sample tests in their study differ little from job knowledge tests, especially if knowledge is broadly defined to include items assessing application of knowledge to the tasks in the job. The motor work samples, from jobs that include such tasks as computer programming and map reading, are

Table 6
Other Meta-Analysis Studies

Study/predictor	Criterion	Mean validity	No. correlations
Hunter (1980c)			
Age	Training	−.01	90
	Proficiency	−.01	425
	Overall	−.01	515
Education	Training	.20	90
	Proficiency	.10	425
	Overall	.12	515
Experience	Training	.01	90
	Proficiency	.18	425
	Overall	.15	515
Hunter (1982)			
Content valid job knowledge test	Work sample	.78	11
	Performance ratings	.48	10
Asher & Sciarrino (1974)			
Verbal work sample	Training	.55	
	Proficiency	.45	30 verbal
Motor work sample	Training	.45	
	Proficiency	.62	30 motor
Cohen et al. (1974)			
Assessment center	Promotion	.63[a]	
	Performance	.43[b]	

Note. For Hunter (1980c), 90 studies used the training criterion, 425 studies used the proficiency criterion; for Hunter (1982), a total of 16 studies were used; for Cohen et al. (1974), a total of 21 studies were used.
[a] Median correlation.
[b] Corrected for attenuation.

Table 7
Meta-Analyses of Three Predictors

Study/predictor	Criterion	No. studies	No. subjects	Mean validity	No. correlations
Kane & Lawler (1978)					
Peer ratings	Promotion	13	6,909	.49	13
	Supervisor ratings	31	8,202	.49	31
	Training success	7	1,406	.36	7
O'Leary (1980)					
College grade point average	Promotion	17	6,014	.21	17
	Supervisor ratings	11	1,089	.11	11
	Tenure	2	181	.05	2
	Training success	3	837	.30	3
Schmidt, Caplan, et al. (1979) and Johnson et al. (1980)					
Traditional t & e rating				.13[a]	65
Behavioral consistency e rating				.49	5

Note. t = training; e = experience.
[a] Corrected for unreliability of supervisor ratings.

harder to describe, but are apparently distinguished by a subjective assessment of the overlap between the test and the job.

The assessment center meta-analysis is now discussed.

New Meta-Analyses

Table 7 presents new meta-analyses by the authors that are little more than reanalyses of studies gathered in three comprehensive narrative reviews. Kane & Lawler (1978) missed few studies done on peer ratings, and O'Leary (1980) was similarly comprehensive in regard to studies using college grade point average. Schmidt, Caplan, et al. (1979) completely reviewed studies of training and experience ratings.

Peer ratings have high validity, but it should be noted that they can be used only in very special contexts. First, the applicant must already be trained for the job. Second, the applicant must work in a context in which his or her performance can be observed by others. Most peer rating studies have been done on military personnel and in police academies for those reasons. The second requirement is well illustrated in a study by Ricciuti (1955), conducted on 324 United States Navy midshipmen. On board the men worked in different sections of the vessel and had only limited opportunity to observe each other at work. In the classroom each person was observed by all his classmates. Because of this fact, even though the behavior observed on summer cruises is more similar to eventual naval duty than behavior during the academic term, the peer ratings for the cruise had an average validity of only .23 (uncorrected), whereas the average validity for peer ratings during academic terms was .30 (uncorrected).

Schmidt, Caplan, et al. (1979) reviewed studies of training and experience ratings. The researchers contrasted two types of procedures: traditional ratings, which consider only amounts of training, education, and experience; and behavioral consistency measures, which try to assess the quality of performance in past work experience. They reported on two meta-analytic studies on traditional training and experience ratings: Mosel (1952) found an average correlation of .09 for 13 jobs, and Molyneaux (1953) found an average correlation of .10 for 52 jobs. If we correct for the unreliability of supervisor ratings, then the average for 65 jobs is .13. Schmidt, Caplan, et al. (1979) reported the results of four studies that used procedures similar to their proposed behavioral consistency method: Primoff (1958), Haynes (undated), and two studies by Acuff (1965). In addition, there has since been a study that used their procedure—Johnson et al. (1980). The meta-analysis of these five studies shows an average validity of .49, which

is far higher than the validity of traditional ratings.

However, the comparison of traditional training and experience ratings with the behavioral consistency method is to some extent unfair. The behavioral consistency method data reviewed here were from situations in which the applicants had already worked in the job for which they were applying. Traditional ratings have often been applied in entry-level situations where the person is to be trained for the job after hiring. Generalized behavioral consistency scales have been constructed and used, but there have been no criterion-related validation studies of such scales to date.

Table 8 presents the new meta-analyses done for this report. Supervisor ratings were corrected for error of measurement. Promotion in most studies meant job level attained (i.e., salary corrected for years with the company or number of promotions), although several studies used sales. Training success refers to training class grades, although most studies used a success–failure criterion. Tenure refers to length of time until termination.

There is a theoretically interesting comparison between training success findings and su-

Table 8
New Meta-Analyses of the Validity of Some Commonly Used Alternatives to Ability Tests

Alternative measure	Average validity	*SD* of validity	No. correlations	Total sample size
		Criterion: supervisor ratings		
Peer ratings	.49	.15	31	8,202
Biodata[a]	.37	.10	12	4,429
Reference checks	.26	.09	10	5,389
College GPA	.11	.00	11	1,089
Interview	.14	.05	10	2,694
Strong Interest Inventory	.10	.11	3	1,789
		Criterion: promotion		
Peer ratings	.49	.06	13	6,909
Biodata[a]	.26	.10	17	9,024
Reference checks	.16	.02	3	415
College GPA	.21	.07	17	6,014
Interview	.08	.00	5	1,744
Strong Interest Inventory	.25	.00	4	603
		Criterion: training success		
Peer ratings	.36	.20	7	1,406
Biodata[a]	.30	.11	11	6,139
Reference checks	.23		1	1,553
College GPA	.30	.16	3	837
Interview	.10	.07	9	3,544
Strong Interest Inventory	.18	.00	2	383
		Criterion: tenure		
Peer ratings[b]				
Biodata[a]	.26	.15	23	10,800
Reference checks	.27	.00	2	2,018
College GPA	.05	.07	2	181
Interview	.03	.00	3	1,925
Strong Interest Inventory	.22	.11	3	3,475

Note. GPA = grade point average.
[a] Only cross validities are reported.
[b] No studies were found.

pervisor rating findings. Table 2 shows that for cognitive ability, the validity is usually higher for training success than for supervisor ratings. This is not generally true for the other predictors in Table 8 (or for psychomotor ability either). This fits with the higher emphasis on learning and remembering in training.

In Table 8, it is interesting to note that across all jobs, the average validity of college GPA in the prediction of promotion was .21 with a standard deviation of .07. A standard deviation of .07 is large in comparison with a mean of .21. A further analysis reveals the reason for this variation. Additional analysis by the authors shows an average validity of GPA for managers of .23 across 13 studies with a total sample size of 5,644. The average validity for people in sales, engineering, and technical work was −.02 across four studies with a sample size of 370. This might mean that the motivational aspects of achieving high college grades are more predictive of motivation in management than in other professions. It might also mean that college grades are considered part of the management evaluation process. That is, it may be that college grades are a criterion contaminate in management promotion. These hypotheses could be tested if there were more studies comparing college grades with supervisor ratings, but only one such study was found for managers and it had a sample size of only 29.

The validity coefficients in Table 8 are not free of sampling error, especially those where the total sample size is below 1,000. For example, the average correlation of .05 for college GPA as a predictor of tenure is based on a total sample size of only 181. The 95% confidence interval is .05 + .15. However, the average correlations for supervisor ratings are all based on total sample sizes of 1,789 or more, with an average of 4,753.

Problems Associated With Two Widely Used Alternative Predictors

Biographical data. For entry-level jobs, biodata predictors have been known to have validity second in rank to that of measures of ability. The average validity of biodata measures is .37 in comparison with .53 for ability measures (with supervisor ratings as the job performance measure). Thus, to substitute biodata for ability would mean a loss of 30% of the saving gained by using ability tests rather than random selection, whereas other predictors would cause larger losses. However, the operational validity of biodata instruments may be lower than the research validity for several reasons.

Biodata predictors are not fixed in their content, but rather constitute a method for constructing fixed predictors. The data reported in the literature are based on specialized biodata instruments; each study used a keyed biographical form where the key is specific to the organization in which the key is tested. There is evidence to suggest that biodata keys are not transportable.

Biodata keys also appear to be specific to the criterion measure used to develop the key. For example, Tucker, Cline, and Schmitt (1967) developed separate keys for supervisor ratings and for tenure that were both checked in the same population of pharmaceutical scientists. A cross validation (as described below) showed that the ratings key predicted ratings (r = .32), and that the tenure key predicted tenure (r = .50). However, the cross correlations between different criteria were different—they were both negative (r = −.07 for the ratings key predicting tenure, and r = −.09 for the tenure key predicting ratings). That is, in their study the key predicting one criterion is actually negatively predictive of the other.

Biodata keys often suffer decay in validity over time. Schuh (1967) reviewed biodata studies in which follow-up validity studies had been done. Within these studies, the mean tryout validity is .66, but in successive follow-up validity studies this value drops from .66 to .52 to .36 to .07. Brown (1978) claimed to have found an organization in which this decay did not occur. However, Brown evaluated his data on the basis of statistical significance rather than on the basis of validity. Because the data were collected from a sample of 14,000, even small correlations were highly significant. Thus, all his correlations registered significant regardless of how small they were and how much they differed in magnitude. His data showed the validity of a key developed for insurance salespersons in 1933 on follow-up studies conducted in 1939 and 1970. If we study validity rather than significance level, the validities were .22 in 1939 and .08 in 1970;

this shows the same decay as found in other studies.

There is a severe feasibility problem with biodata. The empirical process of developing biodata keys is subject to massive capitalization on chance. That is, evidence may be built from what is only a chance relation. The typical key is developed on a sample of persons for whom both biodata and performance ratings are known. Each biographical item is separately correlated with the criterion, and the best items are selected to make up the key. However, on a small sample this procedure picks up not only items that genuinely predict the criterion for all persons, but also items that by chance work only for the particular sample studied. Thus, the correlation between the keyed items and the criterion will be much higher on the sample used to develop the key than on the subsequent applicant population. This statistical error can be eliminated by using a cross-validation research design, in which the sample of data is broken into two subsamples. The key is then developed on one sample, but the correlation that estimates the validity for the key is computed on the other sample. Cross-validity coefficients are not subject to capitalization on chance. Only cross validities are reported in Table 8.

Although the cross-validation research design eliminates the bias in the estimate of the validity that results from capitalization on chance, it does not eliminate the faulty item selection that results from capitalization on chance. Sampling error hurts item selection in two ways: Good items are missed, and poor items are included. For example, suppose that we start with an item pool of 130 items made up of 30 good items, with a population correlation of .13, and 100 bad items, with a population correlation of zero. Let the average interitem correlation be .20. Then, using a sample of 100, we would on the average select 11 of the good items and 5 of the bad items, resulting in a test with a validity of .18. Using a sample of 400, we would select 25 of the good items and 5 of the bad items, resulting in a test with a validity of .23. That is, the validity of the biodata key increases by 28% when the sample size is increased from 100 to 400.

Thus, the use of biodata requires a large sample of workers, say 400 to 1,000, for the tryout study, and more large samples approximately every 3 years to check for decay in validity. Only very large organizations have any jobs with that many workers, so that the valid use of biodata tends to be restricted to organizations that can form a consortium.

One such consortium deserves special mention, that organized by Richardson, Henry, and Bellows for the development of the Supervisory Profile Record (SPR). The latest manual shows that the data base for the SPR has grown to include input from 39 organizations. The validity of the SPR for predicting supervisor ratings is .40, with no variation over time, and little variation across organizations once sampling error is controlled. These two phenomena may be related. The cross-organization development of the key may tend to eliminate many of the idiosyncratic and trivial associations between items and performance ratings. As a result, the same key applies to new organizations and also to new supervisors over time.

Richardson, Henry, and Bellows are also validating a similar instrument called the Management Profile Record (MPR). This instrument is used with managers above the level of first-line supervisor. It has an average validity of .40 across 11 organizations, with a total sample size of 9,826, for the prediction of job level attained. The correlation is higher for groups with more than 5 years of experience.

Two problems might pose legal threats to the use of biographical items to predict work behavior: Some items might be challenged for invasion of privacy, and some might be challenged as indirect indicators of race or sex (see Hunter & Schmidt, 1976, pp. 1067–1068).

Assessment centers. Assessment centers have been cited for high validity in the prediction of managerial success (Huck, 1973), but a detracting element was found by Klimoski and Strickland (1977). They noted that assessment centers have very high validity for predicting promotion but only a moderate validity for predicting actual performance. Cohen et al. (1974) reviewed 21 studies and found a median correlation of .63 predicting potential or promotion but only .33 (.43 corrected for attenuation) predicting supervisor ratings of performance. Klimoski and Strickland interpreted this to mean that assessment centers were acting as policy-capturing devices

which are sensitive to the personal mannerisms that top management tends to use in promotion. To the extent that these mannerisms are unrelated to actual performance, the high correlation between assessment centers and promotion represents a shared error in the stereotype of a good manager.

In a follow-up study, Klimoski and Strickland (1981) gathered data on 140 managers. Their predictors were preassessment supervisor ratings of promotion potential, supervisor ratings of performance, and the assessment center rating. Three years later the criterion data were management grade level attained, supervisor rating of potential, and supervisor rating of performance. The assessment center rating predicted grade level with a correlation of .34 and predicted future potential rating with a correlation of .37, but predicted future performance with a correlation of −.02. Preassessment supervisor ratings of potential predicted future ratings of potential with a correlation of .51 and grade level with a correlation of .14, but predicted future performance ratings with a correlation of only .08. Preassessment performance ratings predicted performance ratings 3 years later with a correlation of .38, but predicted ratings of potential and grade level with correlations of .10 and .06. Thus, assessment center ratings and supervisor ratings of potential predict each other quite highly, but both are much poorer predictors of performance. Klimoski and Strickland interpreted these data to mean that managerial performance could be better predicted by alternatives to assessment centers, such as current performance ratings.

Klimoski and Strickland's (1981) correlations of .37 and .02 between assessment center ratings and ratings of potential and performance are lower, though in the same direction, than the average correlations found by Cohen et al. (1974), which were .63 and .33, respectively. (Note that the sample size in the Klimoski and Strickland study was only 140.) One might argue that in the Klimoski and Strickland study there was a big change in content over the 3-year period, which made future performance unpredictable. However, current job performance ratings predict future job performance ratings in the Klimoski and Strickland study with a correlation of .38. That is, current ratings of job performance predict future ratings of job performance as well as the assessment center predicts future potential ratings.

There appear to be two alternative hypotheses to explain the data on assessment centers. First, it may be that supervisors actually generate better ratings of performance when asked to judge potential for promotion than when asked to judge current performance. Alternatively, it may be that assessment centers, supervisors, and upper-level managers share similar stereotypes of the good manager. To the extent that such stereotypes are valid, the assessment center will predict later performance, as in the Cohen et al. (1974) study showing an average correlation of .43 in predicting later supervisor performance ratings. The higher correlations for ratings of potential against promotion record would then measure the extent to which they share erroneous as well as valid stereotypic attributes.

Comparison of Predictors

If predictors are to be compared, the criterion for job performance must be the same for all. This necessity inexorably leads to the choice of supervisor ratings (with correction for error of measurement) as the criterion because there are prediction studies for supervisor ratings for all predictors. The following comparisons are all for studies using supervisor ratings as the criterion. The comparisons are made separately for two sets of predictors—those that can be used for entry-level jobs where training will follow hiring, and those used for promotion or certification. Table 9 shows the mean validity coefficient across all jobs for 11 predictors suitable for predicting performance in entry-level jobs. Predictors are arranged in descending order of validity. The only predictor with a mean validity of essentially zero is age. The validities for experience, the interview, training and experience ratings, education, or the Strong Interest Inventory are greater than zero but much lower than the validity for ability. Later analysis shows that even the smallest drop in validity would mean substantial losses in productivity if any of these predictors were to be substituted for ability. In fact, the loss is directly proportional to the difference in validity.

Table 10 presents the average validity across all jobs for the predictors that can be used in

Table 9
Mean Validities and Standard Deviations of Various Predictors for Entry-Level Jobs for Which Training Will Occur After Hiring

Predictor	Mean validity	*SD*	No. correlations	No. subjects
Ability composite	.53	.15	425	32,124
Job tryout	.44		20	
Biographical inventory	.37	.10	12	4,429
Reference check	.26	.09	10	5,389
Experience	.18		425	32,124
Interview	.14	.05	10	2,694
Training and experience ratings	.13		65	
Academic achievement	.11	.00	11	1,089
Education	.10		425	32,124
Interest	.10	.11	3	1,789
Age	−.01		425	32,124

situations where the persons selected will be doing the same or substantially the same jobs as they were doing before promotion. Indeed, five of the six predictors (all but ability) are essentially ratings of one or more aspects or components of current performance. Thus, these predictors are essentially predicting future performance from present or past performance. By a small margin, ability is on the average not the best predictor; the work sample is best. The validity of all these predictors is relatively close in magnitude. The drop from work sample test to assessment center is about the same as the drop from ability to the second best predictor for entry-level jobs (job tryout). Even so, the following economic analysis shows that the differences are not trivial.

Combinations of Predictors

For entry-level hiring, it is clear that ability is the best predictor in all but unusual situations (those in which biodata might predict as well as ability). However, there is an important second question: Can other predictors be used in combination with ability to increase validity? There are two few studies with multiple predictors to answer this question directly with meta-analysis. However, there are mathematical formulas that place severe limits on the extent to which the currently known alternatives to ability can improve prediction. This section discusses two such limits.

First, there is a mathematical inequality that shows that the increase in validity for an alternate predictor used in combination with ability is at most the square of its validity. Second, we note that if a second predictor is used with ability, then it will increase validity only if the second predictor is given the small weight that it deserves.

Consider first the maximum extent to which the multiple correlation for ability and an alternate predictor used together can exceed the validity of ability used alone (i.e., can exceed an average of .53). In the few studies using more than one type of predictor, the correlations between alternative predictors tend to be positive. More importantly, the beta weights are all positive. That is, there is no evidence of suppressor effects in predictor combinations. If we can assume that there are no suppressor effects, then we can derive an extremely useful inequality relating the multiple correlation R to the validity of the better predictor (which we will arbitrarily label Variable 1), that is, a comparison between R and r_1. Assuming R is the multiple correlation, and r_1 and r_2 are the zero-order correlations for the individual predictors, then if there are no suppressor effects,

$$R \leq \sqrt{r_1^2 + r_2^2} = r_1 \sqrt{1 + \frac{r_2^2}{r_1^2}}.$$

We can then use Taylor's series to obtain an inequality in which the square root is eliminated:

$$\sqrt{1 + \frac{r_2^2}{r_1^2}} \leq 1 + \tfrac{1}{2} \frac{r_2^2}{r_1^2}.$$

Table 10
Mean Validities and Standard Deviations of Predictors to be Used for Promotion or Certification, Where Current Performance on the Job is the Basis for Selection

Predictor	Mean validity	*SD*	No. correlations	No. subjects
Work sample test	.54			
Ability composite	.53	.15	425	32,124
Peer ratings	.49	.15	31	8,202
Behavioral consistency experience ratings	.49	.08	5	
Job knowledge test	.48	.08	10	3,078
Assessment center	.43			

Combining these inequalities yields

$$R \leq r_1 + \frac{r_2^2}{2r_1}.$$

We denote the validity of the ability composite as r_A and the validity of the alternate predictor as r_X. Then the validity of the composite is at most

$$R \leq r_A + \left(\frac{1}{2r_A}\right) r_X^2.$$

For the average case in which $r_A = .53$, this means that the composite validity is at most

$$R \leq r_A + .94r_X^2 \leq r_A + r_X^2.$$

That is, the increase in validity due to the alternate predictor is at most the square of its validity. For example, the validity of the interview for an average job is .14. If the interview were used in a multiple regression equation with ability, the increase in validity would be at most from .53 to .55.

Second, if an additional predictor is used with ability, it will increase validity only if the second predictor is given the small weight it deserves. Many companies now use several predictors for selection. However, the predictors are not being combined in accordance with the actual validities of the predictors but are weighted equally. Under those conditions the validity of the total procedure can be lower than the validity of the best single predictor. For example, consider an average job in which the interview is given equal weight with the appropriate ability composite score, instead of being given its beta weight. In that case, although the validity of ability alone would be .53, the validity of the inappropriate combination of ability and interview would be at most .47.

Economic Benefits of Various Selection Strategies

Differences in validity among predictors only become meaningful when they are translated into concrete terms. The practical result of comparing the validities of predictors is a number of possible selection strategies. Then the benefit to be gained from each strategy may be calculated. This section presents an analysis of the dollar value of the work-force productivity that results from various selection strategies.

Three strategies are now in use for selection based on ability: hiring from the top down, hiring from the top down within racial or ethnic groups using quotas, and hiring at random after elimination of those with very low ability scores. There is also the possibility of replacing ability measures by other predictors. These strategies can be compared using existing data. The comparison is presented here using examples in which the federal government is the employer; but equations are presented for doing the same utility analysis in any organization.

An analysis of the results of using alternative predictors in combination cannot be made because current data on correlations between predictors are too sparse. However, in the case of the weaker predictors only severely limited improvement is possible.

Computation of Utility

Any procedure that predicts job performance can be used to select candidates for hiring or promotion. The work force selected by a valid procedure will be more productive than a work force selected by a less valid procedure or selected randomly. The extent of

improvement in productivity depends on the extent of accuracy in prediction. If two predictors differ in validity, then the more valid predictor will generate better selection and hence produce a work force with higher productivity. The purpose of this section is to express this difference in productivity in meaningful quantitative terms.

A number of methods have been devised to quantify the value of valid selection (reviewed in Hunter, 1981a). One method is to express increased productivity in dollar terms. The basic equations for doing this were first derived more than 30 years ago (Brogden, 1949; Cronbach & Gleser, 1965), but the knowledge necessary for general application of these equations was not developed until recently (Hunter, 1981a; Hunter & Schmidt, 1982a, 1982b; Schmidt, Hunter, McKenzie, & Muldrow, 1979). The principal equation compares the dollar production of those hired by a particular selection strategy with the dollar production of the same number of workers hired randomly. This number is called the (marginal) utility of the selection procedure. Alternative predictors can be compared by computing the difference in utility of the two predictors. For example, if selection using ability would mean an increased production of $15.61 billion and selection using biodata would mean an increased production of $10.50 billion, then the opportunity cost of substituting biodata for ability in selection would be $5.11 billion.

The utility of a selection strategy depends on the specific context in which the strategy is used. First, a prediction of length of service (tenure) must be considered. The usual baseline for calculating utility is the increase in productivity of people hired in one year. Assume that N persons are to be hired in that year. The gains or losses from a selection decision are realized not just in that year, but over the tenure of each person hired. Moreover, this tenure figure must include not only the time each person spent in their particular jobs, but also the time they spent in subsequent jobs if they were promoted. Persons who can serve in higher jobs are more valuable to the organization than persons who spend their entire tenure in the jobs for which they were initially selected. In the utility equation, we denote the average tenure for those hired as T years.

Second, the effect of selection depends on the dollar value of individual differences in production. This is measured by the standard deviation of performance in dollar terms, denoted here as σ_y.

Third, the effect of selection depends on the extent to which the organization can be selective; the greater the number of potential candidates, the greater the difference between those selected and those rejected. If everyone is hired, then prediction of performance is irrelevant. If only one in a hundred of those who apply is hired, then the organization can hire only those with the highest qualifications. Selectiveness is usually measured by the selection ratio, the percentage of those considered who are selected. In utility equations, the effect of the selection ratio is entered indirectly, as the average predictor score of those selected: The more extreme the selection ratio, the higher the predictor scores of those hired, and hence the higher the average predictor score. In the utility equation, M denotes the average predictor score of those selected, whereas for algebraic simplicity the predictor is expressed in standard score form (i.e., scored so that for applicants the mean predictor score is 0 and the standard deviation is 1). If the predictor scores have a normal distribution, then the number M can be directly calculated from the selection ratio. Conversion values are presented in Table 11.

This completes the list of administrative factors that enter the utility formula: N = the number of persons hired, T = the average tenure of those hired, σ_y = the standard deviation of performance in dollar terms, and M = the expression of the selection ratio as the average predictor score in standard score form.

One psychological measurement factor enters the utility equation: the validity of the predictor, r_{xy}, which is the correlation between the predictor scores and job performance, for the applicant or candidate population. If this validity is to be estimated by the correlation from a criterion-related validation study, then the observed correlation must be corrected for error of measurement in job performance and corrected also for restriction in range. Validity estimates from validity generalization studies incorporate these corrections.

The formula for total utility then computes the gain in productivity due to selection for one year of hiring:

$$U = NTr_{xy}\sigma_y\bar{x}.$$

Table 11
Average Standard Score on the Predictor Variable for Those Selected at a Given Selection Ratio Assuming a Normal Distribution on the Predictor

Selection ratio	Average predictor score M
100	0.00
90	0.20
80	0.35
70	0.50
60	0.64
50	0.80
40	0.97
30	1.17
20	1.40
10	1.76
5	2.08

The standard deviation of job performance in dollars, σ_y, is a difficult figure to estimate. Most accounting procedures consider only total production figures, and totals reflect only mean values; they give no information on individual differences in output. Hunter and Schmidt (1982a) solved this problem by their discovery of an empirical relation between σ_y and annual wage; σ_y is usually at least 40% of annual wage. This empirical baseline has subsequently been explained (Schmidt & Hunter, 1983) in terms of the relation of the value of output to pay, which is typically almost 2 to 1 (because of overhead). Thus, a standard deviation of performance of 40% of wages derives from a standard deviation of 20% of output. This figure can be tested against any study in the literature that reports the mean and standard deviation of performance on a ratio scale. Schmidt and Hunter (1983) recently compiled the results of several dozen such studies, and the 20% figure came out almost on target.

To illustrate, consider Hunter's (1981a) analysis of the federal government as an employer. According to somewhat dated figures from the Bureau of Labor Statistics, as of mid-1980 the average number of people hired in one year is 460,000, the average tenure is 6.52 years, and the average annual wage is $13,598. The average validity of ability measures for the government work force is .55 (slightly higher than the .53 reported for all jobs, because there are fewer low-complexity jobs in government than in the economy as a whole). The federal government can usually hire the top 10% of applicants (i.e., from Table 11, $M = 1.76$). Thus, the productivity gain for one year due to hiring on the basis of ability rather than at random is

$$U = NTr_{xy}\sigma_y\bar{x} = (460{,}000)(6.52)(.55)$$
$$\times (40\% \text{ of } \$13{,}598)(1.76) = \$15.61 \text{ billion.}$$

More recent figures from the Office of Personnel Management show higher tenure and a corresponding lower rate of hiring, but generate essentially the same utility estimate.

The utility formula contains the validity of the predictor r_{xy} as a multiplicative factor. Thus, if one predictor is substituted for another, the utility is changed in direct proportion to the difference in validity. For example, if biodata measures are to be substituted for ability measures, then the utility is reduced in the same proportion as the validities, that is, .37/.55 or .67. The productivity increase over random selection for biodata would be .67(15.61), or $10.50 billion. The loss in substituting biodata for ability measures would be the difference between $15.61 and $10.50 billion—a loss of $5.11 billion per year.

Table 12 presents utility and opportunity cost figures for the federal government for all the predictors suitable for entry-level jobs. The utility estimate for job tryout is quite unrealistic because (a) tryouts cannot be used with jobs that require extensive training, and (b) the cost of the job tryout has not been subtracted from the utility figure given. Thus, for most practical purposes, the next best predictor after ability is biodata, with a cost for substituting this predictor of $5.11 billion per year. The cost of substituting other predictors is even higher.

Table 13 presents utility and opportunity cost figures for the federal government for predictors based on current job performance. Again, because of the higher average complexity of jobs in the government, the validity of ability has been entered in the utility formula as .55 rather than .53. A shift from ability to any of the other predictors means a loss in utility. The opportunity cost for work sample tests ($280 million per year) is unrealistically low because it does not take into account the high cost of developing each work sample test, the high cost of redeveloping the test every time the job changes, or the high cost of administering the work sample test. Because the

range of ability is smaller for predictors based on current performance than for entry-level predictors, the range of opportunity costs is smaller as well.

Productivity and Racial Balance

The best known predictor for entry-level jobs is ability. Therefore, any departure from the optimal selection strategy of hiring from the top down on ability will necessarily lead to loss of productivity. It has been shown that ability tests are fair to minority group members in that they correctly predict job performance, indicating that the differences between racial groups in mean performance on ability tests are real. Therefore, the use of ability tests for selection will inevitably mean a lower hiring rate for minority applicants. Many would argue that racial balance in the work force is so important that the merit principle should be abandoned in favor of affirmative action. We do not debate this ethical position here. (See Hunter & Schmidt, 1976, for a review of such discussions.) However, there are utility implications for hiring on a basis other than merit, which is considered here.

There are two strategies that can be used which base hiring on ability but adjust hiring rates in the direction of affirmative action: using quotas and using ability tests with very low cutoff scores. The hiring strategy that meets affirmative action goals with the least loss in productivity is hiring on the basis of ability with quotas. For example, if an organization is to hire 10% of all applicants, it will hire from the top down on ability within each group separately until it has hired the top 10% of white applicants, the top 10% of black ap-

Table 12
Utility of Alternative Predictors That can be Used for Selection in Entry-Level Jobs Compared With Utility of Ability as a Predictor

Predictor/strategy	Utility[a]	Opportunity costs[a]
Ability composite	15.61	
Ability used with quotas	14.83	−0.78
Job tryout	12.49[b]	−3.12
Biographical inventory	10.50	−5.11
Reference check	7.38	−8.23
Experience	5.11	−10.50
Interview	3.97	−11.64
Training and experience rating	3.69	−11.92
Academic achievement (college)	3.12	−12.49
Education	2.84	−12.77
Interest	2.84	−12.77
Ability with low cutoff score[c]	2.50	−13.11
Age	−0.28	−15.89

Note. Economic benefits or losses in productivity are shown for 1 year of hiring, with the federal government as the employer, if all personnel decisions are based on the use of a single predictor. Utility indicates amount saved over random selection; opportunity costs indicate amount lost if a predictor other than ability is used. The use of ability with quotas and ability as a screen (i.e., the use of ability measures with very low cutoff scores) are shown as if they were alternative predictors rather than alternative strategies for using ability as a predictor.

[a] In billions of dollars.

[b] Does not include the cost of transporting applicant to workplace, paying for training and tryout period, or social cost of firing some applicants after they have been on the job.

[c] Cutoff score was set to reject only the bottom 20% of applicants.

Table 13
Utility of Predictors That can be Used for Promotion or Certification Compared With the Utility of Ability as a Predictor

Predictor/strategy	Utility[a]	Opportunity costs[a]
Ability: hiring top down	15.61	
Work sample test	15.33	−.28
Ability: using quotas	14.83	−.78
Peer ratings	13.91	−1.70
Behavioral consistency experience ratings	13.91	−1.70
Job knowledge test	13.62	−1.99
Assessment center	12.20	−3.41
Ability with low cutoff score[b]	2.50	−13.11

Note. Economic benefits or losses in productivity are shown for 1 year of hiring or promoting, with the federal government as the employer, if all personnel decisions are based on the use of a single predictor. Utility indicates amount saved over random selection; opportunity costs indicate amount lost if a predictor other than ability is used. The use of ability with quotas and the use of ability as a screen (i.e., the use of ability measures with very low cutoff scores) are shown as if they were alternative predictors rather than alternative strategies for using ability as a predictor.

[a] In billions of dollars.

[b] Cutoff score was set to reject only the bottom 20% of applicants.

plicants, the top 10% of Hispanic applicants, and so on.

The other affirmative action strategy is to use ability to select, but to set a very low cutoff score and hire randomly from above that cutoff point. This strategy has two disadvantages in comparison with quotas: It leads to the selection of low-ability applicants from all groups, and it does not completely satisfy affirmative action goals.

The utility implications of affirmative-action strategies can be quantified (Hunter, Schmidt, & Rauschenberger, 1977). Once this had been done for the low-cutoff strategy, the disastrous implications were immediately apparent (Hunter, 1979a, 1981a; Mack, Schmidt, & Hunter, undated). These comparisons were illustrated in Tables 12 and 13. Table 12 shows that for entry-level jobs, using ability with quotas is the second-best strategy, whereas the low-cutoff strategy is the worst predictor except for age, which has an average validity of less than zero. Table 13 shows that for promotion or certification the quota method is third and the low-cutoff method is far poorer than any other.

It is evident that the use of low cutoff scores negates most of the gain derived from the predictive power of ability measures. This method is also vastly inferior to quotas for purposes of racial balance. Quotas guarantee that the hiring rate will be the same in each group, but low cutoff scores do not equalize hiring rates. Consider those jobs in which the discrepancy in ability is greatest, jobs of high complexity for which cognitive ability is the sole predictor. If the cutoff is set to hire 10% of the majority of white people, it will select only 1.1% of the black applicants—a relative minority hiring rate of 11%. However, if the cutoff is set so that 80% of applicants pass, the relative minority hiring rate is still only 52%. The use of a low cutoff score raises the relative hiring rate from 11% to 52%, but this is far short of the 100% that would be produced by the use of quotas.

Comparison of the quota method with the low-cutoff method reveals a stark difference. It would cost the government only $780 million to achieve racial balance using ability with quotas. It would cost $13.11 billion (17 times as much) to use the low-cutoff method, which achieves only a poor approximation of racial balance.

Current interest in predictors other than ability has been spurred primarily by the desire to find an equally valid predictor with much less adverse impact. The alternative predictors for entry-level jobs are all much less valid than are ability measures, and their use implies considerable economic loss. Also, they are not without adverse impact. For example, Schmidt, Greenthal, Hunter, Berner, and Seaton (1977) constructed a work sample test to be used for apprentice machinists in the auto industry. They found racial differences in mean production time as large as the racial differences in mean cognitive ability. The sparse literature on predictors other than ability is reviewed in Reilly and Chao (1982).

Future Research

We have shown that, for entry-level jobs, predictors other than ability have validity so much lower that substitution would mean massive economic loss. It is likely, however, that some of the alternative predictors measure social skills or personality traits that are relevant to job performance and are not assessed by cognitive, perceptual, or psychomotor ability tests. If so, then validity could be increased by using such predictors in conjunction with ability tests. However, the current research base is completely inadequate for setting up such programs, because alternative predictors have been studied in isolation. Only a handful of of studies have considered more than one predictor at a time. In particular, the correlations between alternative predictors are unknown; hence, generalized multiple regression cannot be done.

In the introduction to this article, we noted that adverse impact would be reduced if the validity of prediction could be increased. Validity cannot be increased by replacing ability tests by any of the now known alternative predictors. However, validity could be increased in some jobs by adding the appropriate second predictor and using it with the appropriate weight. If that second predictor has less adverse impact, then the composite selection strategy would have more validity and less adverse impact than the use of ability alone. Unfortunately, there is no research base for suggesting a generally useful strategy of that type. Instead, the use of a second predictor with ability would require a large sample local or consortium-based multivariate validation study.

References

Acuff, R. (1965). *A validity study of an aircraft electrician electrical worker job element examination.* Pasadena, CA: U.S. Naval Laboratories in California, Test Development and Personnel Research Section.

Ash, R. A. (1978, August). *Self assessments of five types of typing ability.* Paper presented at the annual meeting of the American Psychological Association, Montreal.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233–241.

Bass, B. M., & Burger, P. (1979). *Assessment of managers: An international comparison.* New York: Free Press.

Boehm, V. R. (1977). Differential prediction: A methodological artifact? *Journal of Applied Psychology, 62,* 146–154.

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

Brown, S. H. (1978). Long-term validity of a personnel history item scoring procedure. *Journal of Applied Psychology, 63,* 673–676.

Brown, S. H. (1981). Validity generalization in the life insurance industry. *Journal of Applied Psychology, 66,* 664–670.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology, 65,* 543–558.

Callender, J. C., & Osburn, H. G. (1981). Testing the consistency of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results of petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Cohen, B., Moses, J. L., & Byham, W. C. (1974). *The validity of assessment centers: A literature review.* Pittsburgh, PA: Development Dimensions Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology, 62,* 641–644.

Dunnette, M. D. (1972). *Validity study results for jobs relevant to the petroleum refining industry.* Washington, DC: American Petroleum Institute.

Farley, J. A., & Mayfield, E. C. (1976). Peer nominations without peers? *Journal of Applied Psychology, 61,* 109–111.

Fine, S. A. (1955). A structure of worker functions. *Personnel and Guidance Journal, 34,* 66–73.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26,* 461–477.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gordon, L. V. (1973). *Work environment preference schedule manual.* New York: Psychological Corp.

Hannan, R. L. (1979). *Work performance as a function of the interaction of ability, work values, and the perceived environment* (Research Report No. 22). College Park: University of Maryland.

Haynes, E. (undated). *Tryout of job element examination for shipfitter.* Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388–395.

Helm, W. E., Gibson, W. A., & Brogden, H. E. (1957). *An empirical test of shrinkage problems in personnel classification research* (Personnel Board Technical Research Note 84). Arlington, VA: Adjutant General's Office, Personnel Research Branch, U.S. Army.

Huck, J. R. (1973). Assessment centers: A review of the external and internal validities. *Personnel Psychology, 26,* 191–212.

Hunter, J. E. (1979a). *An analysis of validity, differential validity, test fairness, and utility for the Philadelphia Police Officers Selection Examination prepared by the Educational Testing Service.* (Report to the Philadelphia Federal District Court, Alvarez v. City of Philadelphia)

Hunter, J. E. (1979b). *Cumulating results across studies: A critique of factor analysis, canonical correlation, MANOVA, and statistical significance testing.* Paper presented at the annual meeting of the American Psychological Association, New York City.

Hunter, J. E. (1980a). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance.* Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J. E. (1980b). Validity generalization and construct validity. In *Construct validity in psychological measurement: Proceedings of a colloquium on theory and application in education and measurement* (pp. 119–130). Princeton, NJ: Educational Testing Service.

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

Hunter, J. E. (1981a). *The economic benefits of personnel selection using ability tests: A state-of-the-art review including a detailed analysis of the dollar benefit of U.S. Employment Service placements and a critique of the low-cutoff method of test use.* Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J. E. (1981b). *Fairness of the General Aptitude Test Battery (GATB): Ability differences and their impact on minority hiring rates.* Washington, DC: U.S. Employment Service, U.S. Dept. of Labor.

Hunter, J. E. (1981c, April). *False premises underlying the 1978 Uniform Guidelines on Employee Selection Procedures: The myth of test invalidity.* Paper presented to the Personnel Testing Council of Metropolitan Washington, Washington, DC.

Hunter, J. E. (1982, May). *The validity of content valid tests and the basis for ranking.* Paper presented at the International Personnel Management Association Conference, Minneapolis, MN.

Hunter, J. E., & Schmidt, F. L. (1976). A critical analysis of the statistical and ethical implications of five definitions of test fairness. *Psychological Bulletin, 83,* 1053–1071.

Hunter, J. E., & Schmidt, F. L. (1982a). The economic benefits of personnel selection using psychological ability tests. *Industrial Relations, 21,* 293–308.

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

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). *Advanced meta-analysis: Quantitative methods for cumulating research findings across studies.* Beverly Hills, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Pearlman, K. (1982). The history and accuracy of validity generalization equations: A response to the Callender and Osburn reply. *Journal of Applied Psychology, 67,* 853–858.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. *Journal of Applied Psychology, 62,* 245–260.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (in press). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds (Ed.), *Perspectives on bias in mental testing.* New York: Plenum Press.

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438–460.

Johnson, J. C., Guffey, W. L., & Perry, R. A. (1980, July). *When is a T and E rating valid?* Paper presented at the annual meeting of the International Personnel Management Association Assessment Council, Boston, MA.

Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. *Psychological Bulletin, 85,* 555–586.

Katzell, R. A., & Dyer, F. J. (1977). Differential validity revived. *Journal of Applied Psychology, 62,* 137–145.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65,* 507–516.

Klimoski, R. J., & Strickland, W. J. (1977). Assessment centers: Valid or merely prescient? *Personnel Psychology, 30,* 353–361.

Klimoski, R. J., & Strickland, W. J. (1981). *The comparative view of assessment centers.* Unpublished manuscript, Department of Psychology, Ohio State University.

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24,* 247–274.

Levine, E. L., Flory, A. P., & Ash, R. A. (1977). Self-assessment in personnel selection. *Journal of Applied Psychology, 62,* 428–435.

Lilienthal, R. A., & Pearlman, K. (1983). *The validity of Federal selection tests for aid technicians in the health, science, and engineering fields* (OPRD 83-1). Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development. (NTIS No. PB83-202051)

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first year grades in law school. *Applied Psychological Measurement, 5,* 281–289.

Mack, M. J., Schmidt, F. L., & Hunter, J. E. (undated). *Dollar implications of alternative models of selection: A case study of park rangers.* Unpublished manuscript. (Available from F. Schmidt, Office of Personnel Management, Washington, DC 20415)

Molyneaux, J. W. (1953). *An evaluation of unassembled examinations.* Unpublished master's thesis, George Washington University, Washington, DC.

Mosel, J. N. (1952). The validity of rational ratings of experience and training. *Personnel Psychology, 1,* 1–10.

O'Connor, E. J., Wexley, K. N., & Alexander, R. A. (1975). Single-group validity: Fact or fallacy? *Journal of Applied Psychology, 60,* 352–355.

O'Leary, B. S. (1980). *College grade point average as an indicator of occupational success: An update* (PRR-80-23). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center. (NTIS No. PB81-121329)

Pearlman, K. (1982). The Bayesian approach to validity generalization: A systematic examination of the robustness of procedures and conclusions. (Doctoral dissertation, George Washington University, 1982). *Dissertation Abstracts International, 42,* 4960-B.

Pearlman, K., & Schmidt, F. L. (1981, August). Effects of alternate job grouping methods on selection procedure validity. In E. L. Levine (Chair), *Job analysis/job formulas: Current perspectives on research and application.* Symposium conducted at the annual meeting of the American Psychological Association, Los Angeles.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Personnel Research and Development Center. (1981). Alternative selection procedures. *Federal Personnel Manual Bulletin 331-3.* Washington, DC: U.S. Office of Personnel Management.

Primoff, E. S. (1958). *Report on validation of an examination for electrical repairer, McClellan Field, California.* Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Raju, N. S., & Burke, M. J. (in press). Two new procedures for studying validity generalization. *Journal of Applied Psychology.*

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–62.

Ricciuti, H. N. (1955). Ratings of leadership potential at the United States Naval Academy and subsequent officer performance. *Journal of Applied Psychology, 39,* 194–199.

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., Caplan, J. R., Bemis, S. E., Decuir, R., Dunn, L., & Antone, L. (1979). *The behavioral consistency method of unassembled examining* (TM-79-21). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center. (NTIS No. PB80-139942)

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980).

Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643–661.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and employee attitudes. *Personnel Psychology, 30,* 187–197.

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. (1980). The future of criterion-related validity. *Personnel Psychology, 33,* 41–60.

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 the estimate derived from studies of selection procedure utility. *Journal of Applied Psychology, 68,* 407–414.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981a). *Selection procedure validity generalization (transportability) results for three job groups in the petroleum industry.* Washington, DC: American Petroleum Institute.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981b). Validity generalization results for jobs in the petroleum industry. *Journal of Applied Psychology, 66,* 261–273.

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

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences 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., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32,* 257–281.

Schuh, A. J. (1967). The predictability of employee tenure: A review of the literature. *Personnel Psychology, 20,* 133–152.

Sharf, J. C. (1981, April). *Recent developments in the field of industrial and personnel psychology.* Paper presented at the conference sponsored by the Personnel Testing Council of Metropolitan Washington and BNA Systems, Washington, DC.

Thorndike, R. L. (1933). The effect of the interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology, 25,* 543–549.

Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement, 8,* 63–70.

Tucker, M. F., Cline, V. B., & Schmitt, J. R. (1967). Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. *Journal of Applied Psychology, 51,* 131–138.

U.S. Employment Service. (1970). *Manual for the USES General Aptitude Test Battery, Section III: Development.* Washington, DC: U.S. Department of Labor, Manpower Administration.

van Rijn, P., & Payne, S. S. (1980). *Criterion-related validity research base for the D.C. firefighter selection test* (PRR-80-28). Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development. (NTIS No. PB81-122087)

Vineberg, R., & Joyner, J. N. (1982). *Prediction of job performance: Review of military studies.* Alexandria, VA: Human Resources Research Organization.

Received January 11, 1984 ■