Journal of Applied Psychology
1979, Vol. 64, No. 6, 609-626

Impact of Valid Selection Procedures on Work-Force Productivity

Frank L. Schmidt
U.S. Office of Personnel Management
Washington, D.C.
and George Washington University

John E. Hunter
Michigan State University

Robert C. McKenzie and Tressie W. Muldrow
U.S. Office of Personnel Management, Washington, D.C.

Decision theoretic equations were used to estimate the impact of a valid test (the Programmer Aptitude Test; PAT) on productivity if it were used to select new computer programmers for one year in (a) the federal government and (b) the national economy. A newly developed technique was used to estimate the standard deviation of the dollar value of employee job performance, which in the past has been the most difficult and expensive item of required information to estimate. For the federal government and the U.S. economy, separately, results are presented for different selection ratios and for different assumed values for the validity of previously used selection procedures. The impact of the PAT on programmer productivity was substantial for all combinations of assumptions. The results support the conclusion that hundreds of millions of dollars in increased productivity could be realized by increasing the validity of selection decisions in this occupation. Likely similarities between computer programmers and other occupations are discussed. It is concluded that the impact of valid selection procedures on work-force productivity is considerably greater than most personnel psychologists have believed.

Questions concerning the economic and productivity implications of valid selection procedures increasingly have come to the fore in industrial-organizational psychology. Dunnette and Borman's chapter in the *Annual Review of Psychology* (in press) includes, for the first time, a separate section on the utility and productivity implications of selection methods. This development is due at least in part to the emphasis placed on the practical utility of selection procedures in some of the litigation in recent years involving selection tests. Hunter and Schmidt (in press) have contended, on the basis of a review of the empirical literature on the economic utility of selection procedures, that personnel psychologists have typically failed to appreciate the magnitude of productivity gains that result from use of valid selection procedures. The major purpose of this study is to illustrate the productivity (economic utility) implications of a valid selection procedure in the occupation of computer programmer in the federal government and in the economy as a whole. However, to set the stage we first review developments related to selection utility.

History and Development of Selection Utility Models

The evaluation of benefit obtained from selection devices has been a problem of continuing interest in industrial psychology. Most

The opinions expressed herein are those of the authors and do not necessarily reflect official policy of the U. S. Office of Personnel Management.

We would like to thank Lee J. Cronbach and Hubert E. Brogden for their comments on an earlier version of this article.

Requests for reprints should be sent to Frank L. Schmidt, who is at Personnel Research and Development Center, U.S. Office of Personnel Management, 1900 E Street, N.W., Washington, D.C. 20415.

attempts to evaluate benefit have focused on the validity coefficient, and at least five approaches to the interpretation of the validity coefficient have been advanced over the years. The oldest of these is the index of forecasting efficiency, (E): $E = 1 - (1 - r_{xy}^2)^{\frac{1}{2}}$, where r_{xy} is the validity coefficient. This index compares the standard error of job performance scores predicted by means of the test (the standard error of estimate) to the standard error that results when there is no valid information about applicants and one predicts the mean level of performance for everyone (the standard deviation of job performance). The index of forecasting efficiency was heavily emphasized in early texts (Hull, 1928; Kelley, 1923) as the appropriate means for evaluating the value of a selection procedure. This index describes a test correlating .50 with job performance as predicting with a standard deviation of estimate errors only 13% smaller than chance, a very unrealistic and pessimistic interpretation of the test's economic value.

The index of forecasting efficiency was succeeded by the coefficient of determination, which became popular during the 1930s and 1940s. The coefficient of determination is simply the square of the validity coefficient, or r_{xy}^2. This coefficient was referred to as "the proportion of variance in the job performance measure accounted for" by the test. The coefficient of determination describes a test of validity of .50 as "accounting for" 25% of the variance of job performance. Although r_{xy}^2 is still occasionally referred to by selection psychologists—and has surfaced in litigation on personnel tests—the "amount of variance accounted for" has no direct relationship to productivity gains resulting from use of a selection device.

Both E and r_{xy}^2 lead to the conclusion that only tests with relatively high correlation with job performance will have significant practical value. Neither of these interpretations recognizes that the value of a test varies as a function of the parameters of the situation in which it is used. They are general interpretations of the correlation coefficient and have been shown to be inappropriate for interpreting the validity coefficient in selection (Brogden, 1946; Cronbach & Gleser, 1965, p. 31; Curtis & Alf, 1969).

The well-known interpretation developed by Taylor and Russell (1939) goes beyond the validity coefficient itself and takes into account two properties of the selection problem—the selection ratio (the proportion of applicants hired) and the base rate (the percentage of applicants who would be "successful" without use of the test). This model yields a much more realistic interpretation of the value of selection devices. The Taylor–Russell model indicates that even a test with a modest validity can substantially increase the percentage who are successful among those selected when the selection ratio is low. For example, when the base rate is 50% and the selection ratio is .10, a test with validity of only .25 will increase the percentage among the selectees who are successful from 50% to 67%, a gain of 17 additional successful employees per 100 hired.

Although an improvement, the Taylor–Russell approach to determining selection utility does have disadvantages. Foremost among them is the need for a dichotomous criterion. Current employees and new hires must be sorted into an unrealistic 2-point distribution of job performance: "successful" and "unsuccessful" (or "satisfactory"and "unsatisfactory"). As a result, information on levels of performance within each group is lost (Cronbach & Gleser, 1965, pp. 123–124, 138). All those within the "successful" group, for example, are implicitly assumed to be equal in value, whether they perform in an outstanding manner or barely exceed the cutoff. This fact makes it difficult to express utility in units that are comparable across situations.

A second disadvantage of the Taylor–Russell model results from the fact that the decision as to where to draw the line to create the dichotomy in job performance is arbitrary. Objective information on which to base this decision is rarely available, and thus different individuals may draw the line at very different points. This state of affairs creates a problem because the apparent usefulness of the selection procedure depends on where the line is drawn. For example, suppose both the selection ratio and the validity are .50. If the dichotomy is drawn so that 90% of non-test-selected employees are assigned to the successful category, the test raises this figure to 97%—a gain of only seven successful employees per 100 hired, or an 8% increase in the success rate. However, if the dichotomy is drawn so that 50% are considered successful, this same test raises the percentage successful to 67, a gain of 17 successful employees per 100 hired, or a 34% increase in the success rate. Finally, if the line is drawn so that only 10% of employees are considered successful, use of the test raises this figure to 17%. Here the gain is 7 additional successful employees per 100 hired, as it was when 90% were considered successful. But here the percentage increase in the success rate is 70% rather than 8%. Thus the Taylor–Russell approach appears to give different answers to the question of how useful a test is, depending on where the arbitrary dichotomy is drawn.

The next major advance was left to Brogden (1949), who used the principles of linear regression to demonstrate how the selection ratio (SR) and the standard deviation of job performance in dollars (SD_y) affect the economic utility of a selection device. Despite the fact that Brogden's derivations are a landmark in the development of selection utility models, they are straightforward and simple to understand.

Let r_{xy} = the correlation between the test (x) and job performance measured in dollar value. The basic linear model is

$$Y = \beta Z_x + \mu_y + e,$$

where Y = job performance measured in dollar value; β = the linear regression weight on test scores for predicting job performance; Z_x = test performance in standard score form in the applicant group; μ_y = mean job performance (in dollars) of randomly selected employees; and e = error of prediction. This equation applies to the job performance of an individual. The equation that gives the average job performance for the selected (s) group (or for any other subgroup) is

$$E(Y_s) = E(\beta Z_{x_s}) + E(\mu_y) + E(e).$$

Since $E(e) = 0$, and β and μ are constants, this becomes

$$\bar{Y}_s = \beta \bar{Z}_{x_s} + \mu_y.$$

This equation can be further simplified by noting that $\beta = r_{xy}(SD_y/SD_x)$, where SD_y is the standard deviation of job performance measured in dollar value among randomly selected employees. Since $SD_x = 1.00$, $\beta = r_{xy}SD_y$. We thus obtain

$$\bar{Y}_s = r_{xy}SD_y\bar{Z}_{x_s} + \mu_y.$$

This equation gives the *absolute* dollar value of average job performance in the selected group. What is needed is an equation that gives the *increase* in dollar value of average performance that results from using the test. That is, we need an equation for marginal utility. Note that if the test were not used, $\bar{Y}_s$ would be μ_y. That is, mean performance in the selected group is the same as mean performance in a group selected randomly from the applicant pool. Thus the increase due to use of a valid test is $r_{xy}SD_y\bar{Z}_{x_s}$. The equation we want is produced by transposing μ_y to give

$$\bar{Y}_s - \mu_y = r_{xy}SD_y\bar{Z}_{x_s}.$$

The value on the right in the above equation is the difference between mean productivity in the group selected using the test and mean productivity in a group selected without using the test, that is, a group selected randomly. The above equation thus gives mean gain in productivity per selectee resulting from use of the test, that is,

$$\Delta\bar{U}/\text{selectee} = r_{xy}SD_y\bar{Z}_{x_s}, \qquad (1)$$

where $\bar{U}$ is utility and $\Delta\bar{U}$ is marginal utility.

Equation 1 states that the average productivity gain in dollars per person hired is the product of the validity coefficient, the average standard score on the test of those hired, and the SD of job performance in dollars. The value $r_{xy}\bar{Z}_{x_s}$ is the mean standard score on the dollar criterion of those selected, $\bar{Z}_y$. Thus utility per selectee is the mean $\bar{Z}$ score on the criterion of those selected times the standard deviation of the criterion in dollars. The only assumption that Equation 1 makes is that the relation between the test and job performance is linear. If we further assume that the test scores are normally distributed, the mean test score of those selected is ϕ/p, where p = the selection ratio, and ϕ = the ordinate in $N(0, 1)$ at the point of cut corresponding to p. Thus Equation 1 can be written

$$\Delta\bar{U}/\text{selectee} = r_{xy}\phi/pSD_y. \qquad (2)$$

The above equations illustrate the critical role of SD_y and suggest the possibility of situations in which tests of low validity have higher utility than tests of high validity. For example:

	r_{xy}	$\bar{Z}_x$	SD_y	$\Delta\bar{U}/$ selectee
Mid-level job (e.g., systems analyst)	.20	1.00	25,000	\$5,000
Lower-level job (e.g., janitor)	.60	1.00	2,000	\$1,200

The total utility of the test depends on the number of persons hired. The total utility (total productivity) gain resulting from use of the test is simply the mean gain per selectee times the number of people selected, N_s. That is, the total productivity gain is:

$$\Delta U = N_s r_{xy} SD_y \bar{Z}_{x_s}.$$

In this example, the average marginal utilities are \$5,000 and \$1,200. If 10 people were hired, the actual utilities would be \$50,000 and \$12,000, respectively. If 1,000 people were to be hired, then the utilities would be \$500,000 and \$120,000, respectively. Obviously the *total* dollar value of tests is greater for large employers than for small employers. However, this fact can be misleading. On a *percentage* basis it is average gain in utility that counts, and that is what counts to each individual employer.

Equations 1 and 2 clearly illustrate the basis for Brogden's (1946) conclusion that the validity coefficient itself is a direct index of selective efficiency. Brogden (1946) showed that given only the assumption of linearity, the validity coefficient is the proportion of maximum utility attained, where maximum utility is the productivity gain that would result from a perfectly valid test. A test with a validity of .50, for example, can be expected to produce 50% of the gain that would result from a perfect selection device (validity = 1.00) used for the same job and at the same selection ratio. A glance at Equation 1 or Equation 2 verifies this verbal statement. Since the validity coefficient enters the equation as a multiplicative factor, increasing or decreasing the validity by any factor will increase or decrease the utility by the same factor. For example, if we increase validity by a factor of two by raising it from .20 to .40, Equation 2 shows that utility doubles. If we decrease validity by a factor of one half by lowering it from 1.00 to .50, utility is cut in half. Equations 1 and 2 also illustrate the fact that there are limitations on the utility of even a perfectly valid selection device. If the selection ratio is very high, the term ϕ/p (or $\bar{Z}_{x_s}$) approaches zero and even a perfect test has little value. If the selection ratio is 1.00, the perfect test has no value at all. Likewise, as SD_y decreases, the utility of even a perfect test decreases. In a hypothetical world in which SD_y were zero, even a perfect test would have no value.

Brogden (1946) further showed that the validity coefficient could be expressed as the following ratio:

$$r_{xy} = \frac{\bar{Z}_{y(x)} - \bar{Z}_{y(r)}}{\bar{Z}_{y(y)} - \bar{Z}_{y(r)}},$$

where $\bar{Z}_{y(x)}$ = the mean job performance (y) standard score for those selected using the test (x); $\bar{Z}_{y(y)}$ = the mean job performance standard score resulting if selection were on the criterion itself, at the same selection ratio; $\bar{Z}_{y(r)}$ = the mean job performance standard score resulting if selection decisions were made randomly (from among the otherwise screened pool of applicants); and r_{xy} = the validity coefficient. Since $\bar{Z}_{y(r)} = 0$ by definition, the above formula reduces to $\bar{Z}_{y(x)}/\bar{Z}_{y(y)}$. This formulation has implications for the development of new methods of estimating selection procedure validity. If reasonably accurate estimates of both $\bar{Z}_{y(x)}$ and $\bar{Z}_{y(y)}$ can be obtained, validity can be estimated without conducting a traditional validity study.

In Equations 1 and 2, the values for r_{xy} and SD_y should be those that would hold if applicants were hired randomly with respect to test scores. That is, they should be values applicable to the applicant population, the group in which the selection procedure is actually used. Values of r_{xy} and SD_y computed on incumbents will typically be underestimates because of reduced variance among incumbents on both test and job performance measures. Values of r_{xy} computed on incumbents can be corrected for range restriction to produce estimates of the value in the applicant pool (Thorndike, 1949, pp. 169–176). The applicant pool is made up of all who have survived screening on any prior selection hurdles that might be employed, for example, minimum educational requirements or physical examinations.

The correlation between the test and a well-developed measure of job performance (y') provides a good estimate of r_{xy}, the correlation of the test with job performance measured in dollars (productivity). It is a safe assumption that job performance and the value of that performance are at least monotonically related. It is inconceivable that lower performance could have greater dollar value than higher performance. Ordinarily, the relation between y' and y will be not only monotonic but also linear. If there are departures from linearity, the departures will typically be produced by leniency in job performance ratings, which leads to ceiling effects in the measuring instrument. The net effect of such ceiling effects is to make the test's correlation with the measure of job performance smaller than its correlation with actual performance, that is, smaller than its true value, making r_{xy}' an underestimate of r_{xy}. An alternative statement of this effect is that ceiling effects due to leniency produce an artificial nonlinear relation between job performance ratings and the actual dollar value of performance. A nonlinear relation of this form would lead to an underestimation of selection utility because the performance measure underestimates the relative value of very high performers.

Values of r_{xy}' should also be corrected for attenuation due to errors of measurement in the criterion. Random error in the observed measure of job performance causes the test's correlation with that measure to be lower than its correlation with *actual* job performance. Since it is the correlation with actual performance that determines test utility, it is the attenuation-corrected estimate that is needed in the utility formulas. This estimate is simply $r_{xy}'/(r_{y'y'})^{\frac{1}{2}}$, where $r_{y'y'}$ is the reliability of the performance measure. (See Schmidt, Hunter, & Urry, 1976, for further discussion of these points.)

The next major advance in this area came in the form of the monumental work by Cronbach and Gleser (1965), *Psychological Tests and Personnel Decisions*. First published in 1957, this work was republished in 1965 in augmented form. The book consists of detailed and sophisticated application of decision theory principles not only to the single-stage fixed-job selection decisions that we have thus far discussed but also the placement and classification decisions and sequential selection strategies. In these latter areas, many of their derivations were indeed new to the field of personnel testing. Their formulas for utility in the traditional selection setting, however, are, as they note (Cronbach and Gleser, 1965, chap. 4), identical to those of Brogden (1949) except that they formally incorporate cost of testing (information gathering) into the equations. This fact is not obvious.

Brogden, it will be recalled, approached the problem from the point of view of mean gain in utility per selectee. Cronbach and Gleser (1965, chap. 4) derived their initial equation in terms of mean gain per applicant. Their initial formula was (ignoring cost of testing for the moment):

$$\Delta\bar{U}/\text{applicant} = r_{xy}SD_y\phi.$$

All terms are as defined earlier. Multiplying by the number of applicants, N, yields total or overall gain in utility. The Brogden formula for overall utility is

$$\Delta U = N_s\Delta\bar{U}/\text{selectee} = N_s r_{xy}SD_y\phi/p. \quad (3)$$

N_s, it will be recalled, is the number selected. If we note that $p = N_s/N$, that is, the ratio of selectees to applicants, we find that Brogden's equation immediately reduces to the Cronbach–Gleser (1965) equation for total utility:

$$\Delta U = Nr_{xy}SD_y\phi.$$

Role of the Cost of Testing

The previous section ignored the cost of testing, which is small in most testing situations. For example, in a typical job situation, the applicant pool consists of people who walk through the door and ask for a job (i.e., there are no recruiting costs). Hiring is then done on the basis of an application blank and a test that are administered by a trained clerical worker at a cost of $10 or so. If the selection ratio is 10%, then the cost of testing per person hired is $10 for each person hired and $90 for the nine persons rejected in finding the person hired, or $100 altogether. This is negligible in relation to the usual magnitude of utility gains. Furthermore, this $100 is a one-time cost, whereas utility gains continue to accumulate over as many years as the person hired stays with the organization. When cost of testing is included, Equation 2 becomes

$$\Delta\bar{U}/\text{selectee} = r_{xy}SD_y\phi/p - C/p, \qquad (4)$$

where C is the cost of testing one applicant.

Although cost of testing typically has only a trivial impact on selection utility, it is possible to conjure up hypothetical situations in which cost plays a critical role. For example, suppose an employer were recruiting one individual for a sales position that would last only 1 year. Suppose further that the employer decided to base the selection on the results of an assessment center that costs $1,000 per assessee and has a true validity of .40. If the yearly value of SD_y for this job is $10,000 and 10 candidates are assessed, the expected gain in productivity is .4 × $10,000 × 1.758 or $7,032. However, the cost of the assessment center is 10 × 1,000 = $10,000, which is $2,966 greater than the expected productivity gain. That is, under these conditions it would cost more to test 10 persons than would be gained in improved performance. If the employer tested only 5 candidates, then the expected gain in performance would be $5,607, whereas the cost of testing would be $5,000 for an expected gain of $607. In this situation, the optimal number to test is 3 persons. The best person of 3 would have an expected gain in performance of $4,469 with a cost of testing of $3,000, for an expected utility of $1,469.

Relation Between SR and Utility

In most situations, the number to be hired is fixed by organizational needs. If the applicant pool is also fixed, the question of which SR would yield maximum utility becomes academic. The SR is determined by circumstances and is not under the control of the employer. However, employers can often exert some control over the size of the applicant pool by increasing or decreasing recruiting efforts. If this is the case, the question is then how many applicants the employer should test to obtain the needed number of new employees in order to maximize productivity gains from selection.

This question can be answered using a formula given by Cronbach and Gleser (1965, p. 309).

$$\phi - pZ_x = C/(r_{xy}SD_y), \qquad (5)$$

where Z_x is the cutting score on the test in Z-score form. This equation must be solved by iteration. Only one value of the SR (i.e., p) will satisfy this equation, and p will always be less than or equal to .50. The value computed for the optimal SR indicates the number that should be tested in relation to the number to be selected. For example, if the number to be selected is 100 and Equation 5 indicates that the optimal SR is .05, the employer will maximize selection utility by recruiting and testing 2,000 candidates (100/.05 = 2,000). The cost of recruiting additional applicants beyond those available without recruitment efforts must be incorporated into the cost of testing term, C. C then becomes the average cost of recruiting and testing one applicant. The lower the cost of testing and recruiting, the larger the number of applicants it is profitable to test in selecting a given number of new employees. Since the cost of testing is typically low relative to productivity gains from selection, the number tested should typically be large relative to the number selected.

In situations in which the applicant pool is constant, statements about optimal SRs typically do not have practical value, since the SR is not under the control of the employer. Given a fixed applicant pool, $\Delta\bar{U}$/selectee increases as SR ratio decreases if cost of testing is not considered. Brogden (1949) showed that when cost of testing is taken into account and when this cost is unusually high, $\Delta\bar{U}$/selectee will be less at very low SRs than at somewhat higher SRs. If the cost of testing per applicant is unusually high, the cost of testing *per selectee* can become greater at extremely low SRs than $\Delta\bar{U}$/selectee, producing a loss rather than a gain in utility. In practice, however, the combination of extremely high testing costs and extremely low SRs that could lead to negative utilities occurs rarely, if ever. When the applicant pool is fixed, the SR that is optimal for $\Delta\bar{U}$/selectee is not necessarily the optimal SR for total gain in utility. Cronbach and Gleser showed that total utility is always greatest when the SR falls at .50. As SR decreases from .50, $\Delta\bar{U}$/selectee increases until it reaches its maximum, the location of which dependson thecost of testing. But as $\Delta\bar{U}$/selectee increases, the number of selectees N_s is decreasing and the product N_s $\Delta\bar{U}$/selectee, or total utility, is also decreasing. In a fixed applicant pool, total gain is always greatest when 50% are selected and 50% are rejected (Cronbach & Gleser, 1965, pp. 38–40).

Reasons for Failure to Employ Selection Utility Models

Despite the availability since 1949 of the utility equations discussed above, applied differential psychologists have been notably slow in carrying out decision-theoretic utility analyses of selection procedures. In our judgment, the scarcity of work in this area is primarily traceable to three facts. First, many psychologists believe that the utility equations presented above are of no value unless the data exactly fit the linear homoscedastic model and all marginal distributions are normal. Many reject the model in the belief that their data do not perfectly meet the assumptions.

Second, psychologists once believed that validity was situationally specific, that there were subtle differences in the performance requirements of jobs from situation to situation that produced (nontrivial) differences in test validities. If this were true, then the results of a utility analysis conducted in a given setting could not be generalized to apparently identical test–job combinations in new settings. Combined with the belief that utility analyses must include costly cost-accounting applications, it is easy to see why belief in the situational specificity of test validities would lead to reluctance to carry out utility analyses.

Third, it has been extremely difficult in most cases to obtain all the information called for by the equations. The SR and cost of testing can be determined reasonably accurately and at relatively little expense. The item of information that has been most difficult to obtain is the needed estimate of SD_y (Cronbach & Gleser, 1965, p. 121). It has generally been assumed that SD_y can be estimated only by the use of costly and complicated cost-accounting methods. These procedures involve first costing out the dollar value of the job behaviors of each employee (Brogden & Taylor, 1950a) and then computing the standard deviation of these values. In an earlier review (Hunter & Schmidt, in press), we were able to locate only two studies in which cost-accounting procedures were used to estimate SD_y. In this study, we will present an alternative to cost-accounting estimates of SD_y.

We now examine each of these reasons in detail.

Are the Statistical Assumptions Met?

The linear homoscedastic model includes the following three assumptions: (a) linearity, (b) equality of variances of conditional distributions, and (c) normality of conditional distributions. As we have shown above, the basic selection utility equation (Equation 1) depends only on linearity. Equation 2 does assume normality of the test score distribution. However, Brogden (1949) and Cronbach and Gleser (1965) introduced this assumption essentially for derivational convenience; it provides an exact relation between the SR $\bar{Z}_{x_s}$. One need not use the normality-based relation $\phi/p = \bar{Z}_{x_s}$ to compute $\bar{Z}_{x_s}$. The value of $\bar{Z}_{x_s}$ can be computed directly. Thus in the final analysis, linearity is the only required assumption.

To what extent do data in differential psychology fit the linear homoscedastic model? To answer this question, we must of necessity examine sample rather than population data. However, it is only conditions in populations that are of interest; simple data are of interest only as a means of inferring the state of nature in populations. Obviously, the larger the sample used, the more clearly the situation in the sample will reflect that in the population, given that the sample is random. A number of researchers have addressed themselves to the question of the fit of the linear homoscedastic model to data in differential psychology.

Sevier (1957), using Ns from 105 to 250, tested the assumptions of linearity, normality of conditional criterion distributions, and equality of conditional variances. The data were from an education study, with cumulative grade point average being the criterion and high school class rank and various test scores being the predictors. Out of 24 tests of the linearity assumption, only 1 showed a departure significant at the .05 level. Out of eight samples tested for equality of conditional variances, only one showed a departure significant at the .05 level. However, 25 of the 60 tests for normality of the conditional criterion distributions were significant at the .05 level. Violation of this assumption throws interpretations of conditional standard deviations based on normal curve tables into some doubt. However, this statistic typically is not used in practical prediction situations, such as selection or placement. Sevier's study indicates that the assumptions of linearity and equality of conditional variances may be generally tenable.

Ghiselli and Kahneman (1962) examined 60 aptitude variables on one sample of 200 cases and reported that fully 40% of the variables departed significantly from the linear homoscedastic model. Ninety percent of these departures were reported to have held up on cross-validation. Tupes (1964) reanalyzed the Ghiselli and Kahneman data and found that only 20% of the relationships departed from the linear homoscedastic model at the .05 level. He also found that three of the "significant" departures from linearity were probably due to typographical or clerical errors in the data. Later, Ghiselli (1964) accepted and agreed with Tupes's reanalysis of his data. Tupes's findings must be interpreted in light of the fact that the frequency of departure from the linear homoscedastic model expected at the .05 level is in fact much greater than 5%. Tupes carried out two statistical tests on each test-criterion relation: one for linearity and one for equality of conditional variances. Thus the expected proportion of data samples in which at least one test is significant is not .05 but rather a little over .09. If three statistical tests are run at the .05 level—one for linearity, one for normality of conditional distributions, and one for homogeneity of conditional distributions—the expected proportion of data samples in which at least one of these tests is significant is approximately .14 when relations in the parent populations are perfectly linear and homoscedastic.

Tiffin and Vincent (1960) found no significant departures from the bivariate normal model in 15 independent samples (ranging in size from 14 to 157 subjects) of test-criterion data. In each set of data, a chi-square test was used to compare the percentage of employees in the "successful" job performance category in each fifth of the test score distribution to the percentages predicted from the normal bivariate surface (which incorporates the linear homoscedastic model) corresponding to the computed validity coefficient. Surgent (1947) performed a similar analysis on similar data and reported the same findings.

Hawk (1970) reported a major study searching for departures from linearity. The data were drawn from 367 studies conducted between 1950 and 1966 on the General Aptitude Test Battery (GATB) used by the U.S. Department of Labor. A total of 3,303 relations, based on 23,428 individuals, between the nine subtests of the GATB and measures of job performance (typically supervisory ratings) were examined. The frequency of departures from linearity significant at the .05 level was .054. Using the .01 level, the frequency was .012. Frequencies closer to the chance level can hardly be imagined. If any substantial proportion of the relations in the Hawk study had in fact been nonlinear, statistical power to detect this fact would have been high—even if statistical power were low for each of the individual 3,303 relations. For example, suppose statistical power to detect nonlinearity had been as low as .30 in each of the individual tests. Then if 40% of the relations were in fact nonlinear, the expected proportion of significant tests for nonlinearity would have been .30 × .40 + .05, or 17%. If only 20% of the relations were truly nonlinear, the expected proportion significant would have been .30 × .20 + .05, or 11%. If only 10% of the relations were truly nonlinear, the expected proportion significant would have been 9%. The obtained proportion was 5.4%. Thus the Hawk study provides extremely strong evidence against the nonlinearity hypothesis. (For further discussion of statistical power in studies of studies, see Hunter & Schmidt, 1978.)

During his years as technical director of what is now the Army Research Institute for the Behavior and Social Sciences, Brogden and his research associate Lubin spent a considerable amount of time and effort attempting to identify nonlinear test-criterion relationships in large samples of military selection data. Although quadratic and other higher-order nonlinear equations sometimes provided impressive fits to the data in the initial sample, not one of the equations cross-validated successfully in a new sample from the same population. In cross-validation samples, the nonlinear functions were never superior to simple linear functions (Brogden, Note 1).

These findings, taken in toto, indicate that the linear homoscedastic model generally fits the data in this area well. The linearity assumption, the only truly critical assumption, is particularly well supported.

We turn now to the question of normality of marginal distributions. In certain forms (see Equation 2), the Brogden–Cronbach–Gleser utility formulas assume, in addition to linearity, a normal distribution for predictor (test) scores. The Taylor–Russell tables, based on the assumption of a normal bivariate surface, assume normality of total test score distribution also. One obviously relevant question is whether violations of this assumption seriously distort utility estimates. Van Naerssen (1963; cf. also Cronbach & Gleser, 1965) found that they do not. He derived a set of utility equations parallel to the Brogden–Cronbach–Gleser equations, except that they were based on the assumption of a rectangular distribution of test scores. He found that when applied to the same set of empirical data, the two kinds of equation produced highly similar utility estimates (p. 288). Cronbach and Gleser (1965) point out that this finding "makes it possible to generalize over the considerable variety of distributions intermediate between normal and rectangular" (p. 160). Results from the Schmidt and Hoffman (1973) study suggests the same conclusion. In their data neither the predictor nor the criterion scores appeared to be normally distributed. Yet the utility estimates produced by the Taylor–Russell tables were only off marginally: 4.09% at SR = .30 and 11.29% at SR = .50.

Thus it appears that an obsessive concern with statistical assumptions is not justified. This is especially true in light of the fact that for most purposes, there is no need for utility estimates to be accurate down to the last dollar. Approximations are usually adequate for the kinds of decisions that these estimates are used to make (van Naerssen, 1963, p. 282; cf. also Cronbach & Gleser, 1965, p. 139). Alternatives to use of the utility equations will typically be procedures that produce larger errors, or even worse, no utility analyses at all. Faced with these alternatives, errors in the 5%–10% range appear negligible. Further, if overestimation of utility is considered more serious than underestimation, one can always employ conservative estimates of equation parameters (e.g., r_{xy}, SD_y) to virtually guarantee against overestimation of utilities.

Are Test Validities Situationally Specific?

The second reason we postulated for the failure of personnel psychologists to exploit the Brogden–Cronbach–Gleser utility models was belief in the doctrine of situational specificity of validity coefficients. This belief precludes generalization of validities from one setting to another, making criterion-related validity studies, and utility analyses, necessary in each situation. The empirical basis for the principle of situational specificity has been the fact that considerable variability in observed validity coefficients is typically apparent from study to study, even when jobs and tests appear to be similar or essentially identical (Ghiselli, 1966). However, there are a priori grounds for postulating that this variance is due to statistical, measurement, and other artifacts unrelated to the underlying relation between test and job performance. There are at least seven such sources of artifactual variance: (a) differences between studies in criterion reliability, (b) differences between studies in test reliability, (c) differences between studies in range restriction, (d) sampling error (i.e., variance due to $N < \infty$), (e) differences between studies in amount and kind of criterion contamination and deficiency (Brogden & Taylor, 1950b), (f) computational and typographical errors (Wolins, 1962), and (g) slight differences in factor structure between tests of a given type (e.g., arithmetic reasoning tests).

In a purely analytical substudy, Schmidt, Hunter, Pearlman, and Shane (1979) showed that the first four sources alone are capable, under specified and realistic circumstances, of producing as much variation in validities as is typically observed from study to study. They then turned to analyses of empirical data. Using 14 distributions of validity coefficients from the published and unpublished literature for various tests in the occupations of clerical worker and first-line supervisor, they found that artifactual variance sources a through d accounted for an average of 62% of the variance in validity coefficients, with a range from 43% to 87%. Thus there was little remaining variance in which situational moderators could operate. In an earlier study (Schmidt & Hunter, 1977), it was found that sources a, c, and d alone accounted for an average of about 50% of the observed variance in distributions of validity coefficients presented by Ghiselli (1966, p. 29). If one could correct for all seven sources of error variance, one would, in all likelihood, consistently find that the remaining variance was zero or near zero. That is, it is likely that the small amounts of remaining variance in the studies cited here are due to the sources of artifactual variance not corrected for. Thus there is now strong evidence that the observed variation in validities from study to study for similar test–job combinations is artifactual in nature. These findings cast considerable doubt on the situational specificity hypothesis.

Rejection of the situational specificity doctrine obviously opens the way to validity generalization. However, validity generalization is possible in many cases even if the situational specificity hypothesis cannot be definitively rejected. After correcting the mean and variance of the validity distribution for sampling error, for attenuation due to criterion unreliability, and for range restriction (based on average values of both), one may find that a large percentage, say 90%, of all values in the distribution lie above the minimum useful level of validity. In such a case, one can conclude with 90% confidence that true validity is at or above this minimum level in a new situation involving the same test type and job without carrying out a validation study of any kind. Only a job analysis is necessary—to ensure that the job at hand is a member of the class of jobs on which the validity distribution was derived. In Schmidt and Hunter (1977), two of the four validity distributions fell into this category, even though only three sources of artifactual variance could be corrected for. In the later study (Schmidt et al., 1979) in which it was possible to correct for four sources of error variance, 12 of the 14 corrected distributions had 90% or more of validities above levels that would typically be indicative of significant practical utility (cf. Hunter & Schmidt, in press).

These methods and findings indicate that in the future, validity generalization will be possible for a wide variety of test–job combinations. Such a development will do much to encourage the application of decision-theoretic utility estimation tools.

Difficulties in Estimating SD$_y$

The third major reason for neglect of the powerful Brogden–Cronbach–Gleser utility model was the difficulty of estimating SD_y. As noted above, the generally recommended procedure for estimating SD_y uses cost-accounting procedures. Such procedures are supposed to be used to estimate the dollar value of performance of a number of individuals (cf. Brogden & Taylor, 1950a), and the *SD* of these values is then computed. Roche's (1961) dissertation illustrates well the tremendous time and effort such an endeavor entails. This study (summarized in Cronbach and Gleser, 1965, pp. 256–266) was carried out on radial drill operators in a large midwestern plant of a heavy equipment manufacturer. A cost-accounting procedure called *standard costing* was used to determine the contribution of each employee to the profits of the company. The procedure was extremely detailed and complex, involving such considerations as cost estimates for each piece of material machined, direct and indirect labor costs, overhead, and perishable tool usage. There was also a "burden adjustment" for below-standard performance. But despite the complexity and apparent objectivity, Roche was compelled to admit that "many estimates and arbitrary allocations entered into the cost accounting" (Cronbach & Gleser, 1965, p. 263). Cronbach, in commenting on the study after having discussed it with Roche, stated that some of the cost-accounting procedures used were unclear or questionable (Cronbach & Gleser, 1965, pp. 266–267) and that the accountants perhaps did not fully understand the utility estimation problem. Thus even given great effort and expense, cost-accounting procedures may nevertheless lead to a questionable final product.

Recently, we have developed a procedure for obtaining rational estimates of SD_y. This method was used in a pilot study by 62 experienced supervisors of budget analysts to estimate SD_y for that occupation. Supervisors were used as judges because they have the best opportunities to observe actual performance and output differences between employees on a day-to-day basis. The method is based on the following reasoning: If job performance in dollar terms is normally distributed, then the difference between the value to the organization of the products and services produced by the average employee and those produced by an employee at the 85th percentile in performance is equal to SD_y. Budget analyst supervisors were asked to estimate both these values; the final estimate was the average difference across the 62 supervisors. The estimation task presented to the supervisors may appear difficult at first glance, but only 1 out of 62 supervisors objected and stated that he did not think he could make meaningful estimates. Use of a carefully developed questionnaire to obtain the estimates apparently aided significantly; a similar questionnaire was used in the present study and is described in the Method section. The final estimate of SD_y for the budget analyst occupation was \$11,327 per year (SE_M = \$1,120). This estimate is based on incumbents rather than applicants and must therefore be considered to be an underestimate.

As noted earlier, it is generally not critical that estimates of utility be accurate down to the last dollar. Utility estimates are typically used to make decisions about selection procedures, and for this purpose only errors large enough to lead to incorrect decisions are of any consequence. Such errors may be very infrequent. Further, they may be as frequent or more frequent when cost-accounting procedures are used. As we noted above, Roche (1961) found that even in the case of the simple and structured job he studied, the cost accountants were frequently forced to rely on subjective estimates and arbitrary allocations. This is generally true in cost accounting and may become a more severe problem as one moves up in the occupational hierarchy. What objective cost-accounting techniques, for example, can be used to assess the dollar value of an executive's impact on the morale of his or her subordinates? It is the jobs with the largest SD_y values, that is, the jobs for which $\Delta\bar{U}$/selectee is potentially greatest, that are handled least well by cost-accounting methods. Rational estimates—to one degree or another—are virtually unavoidable at the higher job levels.

Our procedure has at least two advantages in this respect. First, the mental standard to be used by the supervisor-judges is the estimated cost to the organization of having an outside consulting firm provide the same products and/or services. In many occupations, this is a relatively concrete standard. Second, the idiosyncratic tendencies, biases, and random errors of individual experts can be controlled by averaging across a large number of judges. In our initial study, the final estimate of SD_y was the average across 62 supervisors. Unless this is an upward or downward bias in the group as a whole, such an average should be fairly accurate. In our example, the standard error of the mean was \$1,120. This means that the interval \$9,480–\$13,175 should contain 90% of such estimates. (One truly bent on being conservative could employ the lower bound of this interval in his or her calculations.)

Methods similar to the one described here have been used successfully by the Decision Analysis Group of the Stanford Research Institute (Howard, Note 2) to scale otherwise unmeasurable but critical variables. Resulting measures have been used in the application of decision-theoretic principles to high-level policy decision making in such areas as nuclear power plant construction, corporate risk policies, investment and expansion programs, and hurricane seeding (Howard, 1966; Howard, Matheson, & North, 1972; Matheson, 1969; Raiffa, 1968). All indications are that the response to the work of this group has been positive; these methods have been judged by high-level decision makers to contribute valuably to improvement of socially and economically important decisions.

In most cases, the alternatives to using a procedure like ours to estimate SD_y are unpalatable. The first alternative is to abandon the idea of a utility analysis. This course of action will typically lead to a gross (implicit) underestimate of the economic value of valid selection procedures. This follows if one accepts our contention (Hunter & Schmidt, in press) that the empirical studies that are available indicate much higher dollar values than psychologists have expected. The second alternative in most situations is use of a less systematized, and probably less accurate, procedure for estimating SD_y. Both of these alternatives can be expected to lead to more erroneous decisions about selection procedures.

The Present Study

The procedure for estimating SD_y described here assumes that dollar outcomes are normally distributed. One purpose of the present study is the evaluation of that assumption.

The present study has three purposes: (a) to illustrate the magnitude of the productivity implications of a valid selection procedure, (b) to demonstrate the application of decision-theoretic utility equations, and (c) to test the assumption that the dollar value of employee productivity is normally distributed.

The major reason for our choice of the job of computer programmer was the remarkably accurate validity estimates for this job that a previous study (Schmidt, Rosenberg, & Hunter, Note 3) had provided. Applying the Schmidt–Hunter (Schmidt et al., 1979) validity generalization model to all available validity data for the Programmer Aptitude Test (PAT; Hughes & McNamara, 1959; McNamara & Hughes, 1961), this study found that the percentage of variance in validity coefficients accounted for in the case of job proficiency criteria for the PAT total score was 94%. This finding effectively refutes the situational specificity hypothesis. The estimated true validity was .76. Thus the evidence is strong that the (multivariate) total PAT score validity is high for predicting performance of computer programmers and that this validity is essentially constant across situations (e.g., different organizations; Schmidt et al., Note 3). Since it is total score that is typically used in selecting programmers, this study concerns itself only with total score validity. Because the PAT is no longer available commercially, testing costs had to be estimated. In this study, we assumed a testing cost of $10 per examinee.

Method

Definition of Relevant Job Group

This study focused on selection of computer programmers at the GS-5 through GS-9 levels. GS-5 is the lowest level in this occupational series. Beyond GS-9, it is unlikely that an aptitude test like the PAT would be used in selection. Applicants for higher level programmer positions are expected (and required) to have considerable developed expertise in programming and are selected on the basis of achievement and experience, rather than directly on aptitude. The vast majority of programmers hired at the GS-9 level are promoted to GS-11 after 1 year. Similarly, all but a minority hired at the GS-5 level advance to GS-7 in 1 year and to GS-9 the following year. Therefore, the SD_y estimates were obtained for the GS-9 through GS-11 levels. Statistical information obtained from the Bureau of Personnel Management Information Systems of the U.S. Office of Personnel Management indicated that the number of programmer incumbents in the federal government at the relevant levels (GS-5 through GS-9) was 4,404 (as of October 31, 1976, the latest date for which figures were available). The total number of computer programmers at all grade levels was 18,498. For 1975–1976, 61.3% of all new hires were at the GS-5 through GS-9 levels. The number of new hires governmentwide in this occupation at these levels was 655 and 565 for calendar years 1975 and 1976, respectively, for an average yearly selection rate of 618. The average tenure of computer programmers hired at GS-5 through GS-9 levels was determined to be 9.69 years.

Data from the 1970 U.S. Census (U.S. Bureau of the Census, 1970) showed that there were 166,556 computer programmers in the U.S. in that year. Because the growth rate has been rapid in this occupation recently, this figure undoubtedly underestimates the current number of programmers. However, it is the most recent estimate available. In any event, the effect of underestimation on the utility results is a conservative one. It was not possible to determine the number of computer programmers that are hired yearly in the U.S. economy. For purposes of this study, it was assumed that the turnover rate was 10% in this occupation and that therefore .10 × 166,556, or 16,655, were hired to replace those who had quit, retired, or died. Extrapolating from the federal to the private sector work force, it was assumed that 61.3% of these new hires were at occupational levels for which the PAT would be appropriate. Thus it was assumed that .613 × 16,655, or 10,210 computer programmers could be hired each year in the U.S. economy using the PAT. In view of the current rapid expansion of this occupation, it is likely that this number is a substantial underestimate.

It was not possible to determine prevailing SRs for computer programmers in the general economy. Because the total yearly number of applicants for this job in the government could not be determined, it was also impossible to estimate the government SR. This information lack is of no real consequence, however, since it is more instructive to examine utilities for a variety of SRs. Utilities were calculated for SRs of .05, .10, .20 . . . .80. The gains in utility or productivity as computed from Equation 4 are those that result when a valid procedure is introduced where previously no procedure or a totally invalid procedure has been used. The assumption that the true validity of the previous procedure is essentially zero may be valid in some cases, but in other situations the PAT would, if introduced, replace a procedure with lower but nonzero true validity. Hence, utilities were calculated assuming previous procedure true validities of .20, .30, .40, and .50, as well as .00.

Estimating SD_y

Estimates of SD_y were provided by experienced supervisors of computer programmers in 10 federal agencies. These supervisors were selected by their own supervisors after consultation with the first author. Participation was voluntary. Of 147 questionnaires distributed, 105 were returned (all in usable form), for a return rate of 71.4%. To test the hypothesis that dollar outcomes are normally distributed, the supervisors were asked to estimate values for the 15th percentile ("low-performing programmers"), the 50th percentile ("average programmers"), and the 85th percentile ("superior programmers"). The resulting data thus provide two estimates of SD_y. If the distribution is approximately normal, these two estimates will not differ substantially in value.

The instructions to the supervisors were as follows:

> The dollar utility estimates we are asking you to make are critical in estimating the relative dollar value to the government of different selection methods. In answering these questions, you will have to make some very ***difficult judgments***. We realize they are difficult and that they are judgments or estimates. You will have to ponder for some time before giving each estimate, and there is probably no way you can be absolutely certain your estimate is accurate when you do reach a decision. But keep in mind three things:
>
> (1) The alternative to estimates of this kind is application of cost accounting procedures to the evaluation of job performance. Such applications are usually prohibitively expensive. And in the end, they produce only imperfect estimates, like this estimation procedure.
>
> (2) Your estimates will be averaged in with those of other supervisors of computer programmers. Thus errors produced by too high and too low estimates will tend to be averaged out, providing more accurate final estimates.
>
> (3) The decisions that must be made about selection methods do not require that all estimates be accurate down to the last dollar. Substantially accurate estimates will lead to the same decisions as perfectly accurate estimates.
>
> Based on your experience with agency programmers, we would like for you to estimate the yearly value to your agency of the products and services produced by the average GS 9-11 computer programmer. Consider the quality and quantity of output typical of the ***average programmer*** and the value of this output. In placing an overall dollar value on this output, it may help to consider what the cost would be of having an outside firm provide these products and services.
>
> Based on my experience, I estimate the value to my agency of the average GS 9-11 computer programmer at ______ dollars per year.
>
> We would now like for you to consider the "***superior***" ***programmer***. Let us define a superior performer as a programmer who is at the 85th percentile. That is, his or her performance is better than that of 85% of his or her fellow GS 9-11 programmers, and only 15% turn in better performances. Consider the quality and quantity of the output typical of the superior programmer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.
>
> Based on my experience, I estimate the value to my agency of a superior GS 9-11 computer programmer to be ______ dollars per year.
>
> Finally, we would like you to consider the "***low performing***" ***computer programmer***. Let us define a low performing programmer as one who is at the 15th percentile. That is, 85% of all GS 9-11 computer programmers turn in performances better than the low performing programmer, and only 15% turn in worse performances. Consider the quality and quantity of the output typical of the low performing programmer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.
>
> Based on my experience, I estimate the value to my agency of the low performing GS 9-11 computer programmer at ______ dollars per year.

The wording of this questionnaire was carefully developed and pretested on a small sample of programmer supervisors and personnel psychologists. None of the programmer supervisors who returned questionnaires in the study reported any difficulty in understanding the questionnaire or in making the estimates.

Computation of Impact on Productivity

Using a modification of Equation 4, utilities that would result from 1 year's use of the PAT for selection of ***new hires*** in the federal government and the economy

Table 1
Estimated Productivity Increase From 1 Year's Use of the Programmer Aptitude Test to Select Computer Programmers in the Federal Government (in Millions of Dollars)

Selection ratio	True validity of previous procedure				
	.00	.20	.30	.40	.50
.05	97.2	71.7	58.9	46.1	33.3
.10	82.8	60.1	50.1	39.2	28.3
.20	66.0	48.6	40.0	31.3	22.6
.30	54.7	40.3	33.1	25.9	18.7
.40	45.6	34.6	27.6	21.6	15.6
.50	37.6	27.7	22.8	17.8	12.9
.60	30.4	22.4	18.4	14.4	10.4
.70	23.4	17.2	14.1	11.1	8.0
.80	16.5	12.2	10.0	7.8	5.6

as a whole were computed for each of the combinations of SR and previous procedure validity given above. When the previous procedure was assumed to have zero validity, its associated testing cost was also assumed to be zero; that is, it was assumed that no procedure was used and that otherwise prescreened applicants were hired randomly. When the previous procedure was assumed to have a nonzero validity, its associated cost was assumed to be the same as that of the PAT, that is, $10 per applicant. As mentioned above, average tenure for government programmers was found to be 9.69 years; in the absence of other information, this tenure figure was also assumed for the private sector. $\Delta\bar{U}$/selectee per year was multiplied by 9.69 to give final $\Delta\bar{U}$/selectee. Cost of testing was charged only to the first year.

Building all of these factors into Equation 4, we obtain the equation actually used in computing the utilities:

$$\Delta U = tN_s(r_1 - r_2)SD_y\phi/p - N_s(C_1 - C_2)/p, \quad (6)$$

where ΔU = the gain in productivity in dollars from using the new selection procedure for 1 year; t = tenure in years of the average selectee, here 9.69; N_s = number selected in a given year (this figure was 618 for the federal government and 10,210 for the U.S. economy); r_1 = validity of the new procedure, here the PAT (r_1 = .76); r_2 = validity of the previous procedure (r_2 ranges from 0 to .50); C_1 = per-applicant cost of the new procedure, here $10; and C_2 = per-applicant cost of previous procedure, here 0 or $10. The terms SD_y, ϕ, and p are as defined previously. The figure for SD_y was the average of the two estimates obtained in this study. Note that although this equation gives the productivity gain that results from substituting *for 1 year* the new (more valid) selection procedure for the previous procedure, these gains are not all realized the first year. They are spread out over the tenure of the new employees.

Results and Discussion

Estimation of Yearly SD$_y$

The two estimates of SD_y were similar. The mean estimated difference in dollar value of yearly job performance between programmers at the 85th and 50th percentiles in job performance was $10,871 ($SE$ = $1,673). The figure for the difference between the 50th and 15th percentiles was $9,955 ($SE$ = $1,035). The difference of $916 is roughly 8% of each of the estimates and is not statistically significant. Thus the hypothesis that computer programmer productivity in dollars is normally distributed cannot be rejected. The distribution appears to be at least approximately normal. The average of these two estimates, $10,413, was the SD_y figure used in the utility calculations below. This figure must be considered an underestimate, since it applies to incumbents rather than to the applicant pool. As can be seen from the two standard errors, supervisors showed better agreement on the productivity difference between "low-performing" and "average programmers" than on the difference between "average" and "superior" programmers.

Impact on Productivity

Table 1 shows the gains in productivity in millions of dollars that would result from 1 year's use of the PAT to select computer programmers in the federal government for different combinations of SR and previous procedure validity. As expected, these gains increase as SR decreases and as the validity of the previous procedure decreases. When SR is .05 and the previous procedure has no validity, use of the PAT for 1 year produces a productivity gain of $97.2 million. At the other extreme, if SR is .80 and the procedure the PAT replaces has a validity of .50, the gain is only $5.6 million. The figures in all cells of Table 1 are large—larger than most industrial-organizational psychologists would, in our judgment, have expected. These figures, of course, are for total utility. Gain per selectee for any cell in Table 1 can be computed by dividing the cell entry by 618, the assumed yearly number of selectees. For example, when SR = .20 and the previous procedure has a validity of .30, the gain per selectee is $64,725. As indicated earlier, the gains shown in Table 1 are produced by 1 year's use of the PAT but are not all realized during the first year; they are spread out over the tenure of the new employees. Per-year gains for any cell in Table 1 can be obtained by dividing the cell entry by 9.69, the average tenure of computer programmers.

Table 2 shows productivity gains for the economy as a whole resulting from use of the PAT or substitution of the PAT for less valid procedures. Table 2 figures are based on the assumed yearly selection of 10,210 computer programmers nationwide. Again, the figures are for the total productivity gain, but gain per selectee can be computed by dividing the cell entry by the number selected. Once mean gain per selectee is obtained, the reader can easily compute total gain for any desired number of selectees. As expected, these figures are considerably larger, exceeding $1 billion in several cells. Although we have no direct evidence on this point, we again judge that the productivity gains are much higher than most industrial-organizational psychologists would have suspected.

In addition to the assumptions of linearity and normality discussed earlier, the productivity gain figures in Tables 1 and 2 are based on two additional assumptions. The first is the assumption that selection proceeds from top-scoring applicants downward until the SR has been reached. That is, these analyses assume that selection procedures are used optimally. Because of the linearity of the relation between test score and job performance, any other usage of a valid test would result in lower mean productivity levels among selectees. For example, if a cutting score were set at a point lower than that corresponding to the SR and if applicants scoring above this minimum score were then selected randomly (or selected using other nonvalid procedures or considerations), productivity gains would be considerably lower than those shown in Tables 1 and 2. (They would, however, typically still be substantial.)

The second additional assumption is that all applicants who are offered jobs accept and are hired. This is often not the case, and the effect of rejection of job offers by applicants is to increase the SR and thus lower the productivity gains from selection. For example, if a SR of .10 would yield the needed number of new employees given no rejections by applicants, then if half of all job offers are rejected, the SR must be increased to .20 to yield the desired number of selectees. If the validity of the previous procedure were zero, Table 1 shows that rejection by applicants would reduce productivity gains from $82.8 to $66.0 million, a reduction of $16.8 million. If the validity of the previous procedure were nonzero, job rejection by applicants would reduce both its utility and the utility of the new test. However, the function is multiplicative and hence the utility of the more valid procedure would be reduced by a greater amount. Therefore, the utility advantage of the more valid procedure over the less valid procedure would be reduced. For example, Table 1 shows that if the validity of the previous procedure were .30, the productivity advantage of the more valid test would be $50.1 million if the needed workers could be hired using a selection ratio of .10. But if half of the applicants rejected job offers, we would have to use a SR of .20, and the advantage of the more valid test would drop by one fifth to $40 million.

Table 2
Estimated Productivity Increase From 1 Year's Use of Programmer Aptitude Test to Select Computer Programmers in U.S. Economy (in Millions of Dollars)

Selection ratio	True validity of previous procedure				
	.00	.20	.30	.40	.50
.05	1,605	1,184	973	761	550
.10	1,367	1,008	828	648	468
.20	1,091	804	661	517	373
.30	903	666	547	428	309
.40	753	555	455	356	257
.50	622	459	376	295	213
.60	501	370	304	238	172
.70	387	285	234	183	132
.80	273	201	165	129	93

Hogarth and Einhorn (1976) have pointed out that utility losses caused by job offer rejection can often be offset in part by additional recruiting efforts that increase the size of the applicant pool and, therefore, restore use of smaller SRs. They present equations that allow one to compute the optimal number of additional applicants to recruit and test and the optimal SR under various combinations of circumstances.

The PAT is no longer available commercially. Originally marketed by Psychological Corporation, it was later distributed by IBM as part of package deals to computer systems purchasers. However, this practice was dropped around 1974, and since then the PAT has not been available to most users (Dyer, Note 4). This fact, however, need create no problems in terms of validity generalization. The validity estimates from Schmidt et al. (Note 3) generalize directly to other tests and subtests with the same factor structure. The three subscales of the PAT are composed of conventional number series, figure analogies, and arithmetic reasoning items. New tests can easily be constructed that correlate 1.00, corrected for attenuation, with the PAT subtests.

The productivity gains shown in Tables 1 and 2 are based on an estimated true validity of .76 for the PAT total score (Schmidt et al., Note 3). For many jobs, alternative selection procedures with known validities of this magnitude may not be available. However, the methods used here would be equally applicable. For example, if the alternate selection procedure has an estimated true validity of .50, Equation 6 can be used in the same way to estimate values comparable to those in Tables 1 and 2. Obviously, in this case, all productivity gain would be smaller and there would be no productivity gain at all from substituting the new procedure for an existing procedure with validity of .50. But work-force productivity will often be optimized by combining the existing procedure and the new procedure to obtain validity higher than either procedure can provide individually. This fact is well known in personnel psychology, and we therefore do not develop it further here.

It should be noted that productivity gains comparable to those shown in Tables 1 and 2 can probably be realized in other occupations, such as clerical work, in which lower SD_y values will be offset by the larger numbers of selectees. Pearlman, Schmidt, and Hunter (in press) present extensive data on the generalizability of validity for a number of different kinds of cognitive measures (constructs) for several job families of clerical work.

There is another way to approach the question of productivity gains resulting from use of valid selection procedures. One can ask what the productivity gain would have been had the entire incumbent population been selected using the more valid procedure. As indicated earlier, the incumbent population of interest in the federal government numbers 18,498. As an example, suppose this population had been selected using a procedure with true validity of .30 using a SR of .20. Then had the PAT been used instead, the productivity gain would have been approximately $1.2 billion [9.69 × 18,498 × (.76 − .30) × 10,413 × .28/.20]. Expanding this example to the economy as a whole, the productivity gain that would have resulted is $10.78 billion.

Obviously, there are many other such examples that can be worked out, and we encourage readers to ask their own questions and derive their own answers. However, virtually regardless of the question, the answer always seems to include the conclusion that it does make a difference—an important, practical difference—how people are selected. We conclude that the implications of valid selection procedures for work-force productivity are much greater than most of us have realized in the past.

Finally, we note by way of a necessary caution that productivity gains in individual jobs from improved selection cannot be extrapolated in a simple way to productivity gains in the composite of all jobs making up the national economy. To illustrate, if the potential gain economywide in the computer programmer occupation is $10.78 billion and if there are N jobs in the economy, the gain to be expected from use of improved selection procedures in all N jobs will not in general be as great as N times $10.78 billion. Since the total talent pool is not unlimited, gains due to selection in one job are partially offset by losses in other jobs. The size of the net gain for the economy depends on such factors as the number of jobs, the correlation between jobs of predicted success composites ($\hat{y}$s), and differences between jobs in SD_y. Nevertheless, potential net gains for the economy as a whole are large. The impact of selection procedures on the economy as a whole is explored in detail in Hunter and Schmidt (in press).

Reference Notes

1. Brogden, H. E. Personal communication, November 1967.
2. Howard, R. A. *Decision analysis: Applied decision theory.* Paper presented at the Fourth Internationa\| Conference on Operational Research, Boston, 1966.
3. Schmidt, F. L., Rosenberg, I. G., & Hunter, J. E. *Application of the Schmidt–Hunter validity generalization model to computer programmers.* Personnel Research and Development Center, U.S. Civil Service Commission, Washington, D.C., 1978.
4. Dyer, P. Personal communication, April 20, 1978.

References

Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, 1946, *37*, 65–76.

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

Brogden, H. E., & Taylor, E. K. The dollar criterion: Applying the cost accounting concept to criterion construction. *Personnel Psychology*, 1950, *3*, 133–154. (a)

Brogden, H. E., & Taylor, E. K. A theory and classification of criterion bias. *Educational and Psychological Measurement*, 1950, *10*, 159–186. (b)

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

Curtis, E. W., & Alf, E. F. Validity, predictive efficiency, and practical significance of selection tests. *Journal of Applied Psychology*, 1969, *53*, 327–337.

Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. *Annual Review of Psychology*, 1979, in press.

Ghiselli, E. E. Dr. Ghiselli's comments on Dr. Tupes' note. *Personnel Psychology*, 1964, *17*, 61–63.

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

Ghiselli, E. E., & Kahneman, D. Validity and nonlinear heteroscedastic models. *Personnel Psychology*, 1962, *15*, 1–11.

Hawk, J. Linearity of criterion-GATB aptitude relationships. *Measurement and Evaluation in Guidance*, 1970, *2*, 249–251.

Hogarth, R. M., & Einhorn, H. J. Optimal strategies for personnel selection when candidates can reject offers. *Journal of Business*, 1976, *49*, 478–495.

Howard, R. A. (Ed.). *Proceedings of the Fourth International Conference on Operational Research.* New York: Wiley, 1966.

Howard, R. A., Matheson, J. E., & North, D. W. The decision to seed hurricanes. *Science*, 1972, *176*, 1191–1202.

Hughes, J. L., & McNamara, W. J. *Manual for the revised Programmer Aptitude Test.* New York: Psychological Corporation, 1959.

Hull, C. L. *Aptitude testing.* Yonkers, N.Y.: World Book, 1928.

Hunter, J. E., & Schmidt, F. L. Differential and single group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology*, 1978, *63*, 1–11.

Hunter, J. E., & Schmidt, F. L. Fitting people to jobs: The impact of personnel selection on national productivity. In E. A. Fleishman (Ed.), *Human performance and productivity*, in press.

Kelley, T. L. *Statistical method.* New York: Macmillan, 1923.

Matheson, J. E. Decision analysis practice: Examples and insights. In, *Proceedings of the Fifth International Conference on Operational Research* (*OR 69*). London: Tavistock, 1969.

McNamara, W. J., & Hughes, J. L. A review of research on the selection of computer programmers. *Personnel Psychology*, 1961, *14*, 39–51.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. Test of a new model of validity generalization: Results for job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology*, in press.

Raiffa, H. Decision analysis. In, *Introductory lectures on choices under uncertainty.* Reading, Mass.: Addison-Wesley, 1968.

Roche, U. J. The Cronbach-Gleser utility function in fixed treatment employee selection (Doctoral dissertation, Southern Illinois University, 1961). *Dissertation Abstracts International*, 1961–62, *22*, 4413. (University Microfilms No. 62-1570). (Portions reproduced in L. J. Cronbach & G. C. Gleser (Eds.), *Psychological tests and personnel decisions.* Urbana: University of Illinois Press, 1965.)

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

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

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

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validity studies. *Journal of Applied Psychology*, 1976, *61*, 473–485.

Sevier, F. A. C. Testing the assumptions underlying multiple regression. *Journal of Experimental Education*, 1957, *25*, 323–330.

Sűrgent, L. V. The use of aptitude tests in the selection of radio tube mounters. *Psychological Monographs*, 1947, *61*(2, Whole No. 283).

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 1939, *23*, 565–578.

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

Tiffin, J., & Vincent, N. L. Comparison of empirical and theoretical expectancies. *Personnel Psychology*, 1960, *13*, 59–64.

Tupes, E. C. A note on "Validity and non-linear heteroscedastic models." *Personnel Psychology*, 1964, *17*, 59–61.

U.S. Bureau of the Census. *Census of population: 1970. Subject reports* (Final Rep. PC (2)-7A). Washington, D.C.: Author, 1970.

van Naerssen, R. F. *Selectie van chauffeurs*. Groningen, The Netherlands: Wolters, 1963. (Portions reprinted in L. J. Cronbach & G. C. Gleser [Eds.; J. Wassing, trans.] *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1965.)

Wolins, L. Responsibility for raw data. *American Psychologist*, 1962, *17*, 657–658.

Received December 27, 1978 ■

Manuscripts Accepted for Publication

Role Theory, Attitudinal Constructs, and Actual Performance: A Measurement Issue. Eric N. Berkowitz (College of Business Administration and Graduate School of Business Administration, Business Administration Building, 271 19th Avenue South, University of Minnesota, Minneapolis, Minnesota 55455).

Measuring the Relative Importance of Utilitarian and Egalitarian Values: A Study of Individual Differences About Fair Distribution. John Rohrbaugh (Graduate School of Public Affairs, The State University of New York at Albany, 1400 Washington Avenue, Albany, New York 12222), Gary McClelland, and Robert Quinn.

Interview Behaviors of Mentally Retarded Adults as Predictors of Employability. Carol K. Sigelman (Texas Tech University, Research and Training Center in Mental Retardation, P. O. Box 4510, Lubbock, Texas 79409), Susan F. Elias, and Pamela Danker-Brown.

Long-Term Auditory Memory: Speaker Identification. Howard Saslove and A. Daniel Yarmey (Department of Psychology, University of Guelph, Guelph, Ontario, Canada N1G 2W1).

Effects of Student Participation in Classroom Decision Making on Attitudes, Peer Interaction, Motivation, and Learning. Fredrick D. Richter and Dean Tjosvold (Department of Economics and Commerce, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6).

Machiavellianism and Leadership. Amos Drory (Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beersheba, Israel) and Uri M. Gluskinos.

Effects of Noise on Performance on Embedded Figures Tasks. Andrew P. Smith and Donald E. Broadbent (Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, England).

Leader–Group Interactions: A Longitudinal Field Investigation. Charles N. Greene (Indiana University, School of Business, 10th and Fee Lane, Bloomington, Indiana 47405) and Chester A. Schriesheim.

Redundancy and Dimensionality as Determinants of Data Analytic Strategies in Multivariate Analysis of Variance. Paul E. Spector (Northside Community Mental Health Center, Inc., 13301 North 30th Street, Tampa, Florida 33612).

Testing Mintzberg's Managerial Roles Classification Using an In-Basket Simulation. Zur Shapira (School of Business, Hebrew University, Jerusalem, Israel) and Roger L. M. Dunbar.

Evaluation of Feedback Sources as a Function of Role and Organizational Level. Martin M. Greller (Rohrer, Hibler & Replogle, Inc., Suite 610, 10 Rockefeller Plaza, New York, New York 10020).

Comprehending Spatial Information: The Relative Efficiency of Different Methods of Presenting Information About Bus Routes. D. J. Bartram (Department of Psychology, University of Hull, Hull, North Humberside HU6 7RY, England).

Effects of Rater Training: Creating New Response Sets and Decreasing Accuracy. H. John Bernardin (Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061) and Earl C. Pence.

Individual Differences in Electrodermal Lability and the Detection of Information and Deception. William M. Waid (Unit for Experimental Psychiatry, 111 North 49th Street, Philadelphia, Pennsylvania 19139) and Martin T. Orne.

(*Continued on page 634*)