Psychology, Public Policy, and Law
1996, Vol. 2, No. 2, 363–376

1076-8971/96/$3.00

SEEING THE FOREST FROM THE TREES: When Predicting the Behavior or Status of Groups, Correlate Means

David Lubinski
Iowa State University

Lloyd G. Humphreys
University of Illinois, Urbana-Champaign

When measures of individual differences are used to predict group performance, the reporting of correlations computed on samples of individuals invites misinterpretation and dismissal of the data. In contrast, if regression equations, in which the correlations required are computed on bivariate means, as are the distribution statistics, it is difficult to underappreciate or lightly dismiss the utility of psychological predictors. Given sufficient sample size and linearity of regression, this technique produces cross-validated regression equations that forecast criterion means with almost perfect accuracy. This level of accuracy is provided by correlations approaching unity between bivariate samples of predictor and criterion means, and this holds true regardless of the magnitude of the "simple" correlation (e.g., r_{xy} = .20, or r_{xy} = .80). We illustrate this technique empirically using a measure of general intelligence as the predictor and other measures of individual differences and socioeconomic status as criteria. In addition to theoretical applications pertaining to group trends, this methodology also has implications for applied problems aimed at developing policy in education, medical, and psychological clinics, business, industry, the military, and other domains of public welfare. Linkages between this approach and epidemiological research reinforce its utility as a tool for making decisions about policy.

The initial response to a correlation coefficient by a large number of psychologists, as well as other professionals who have had training in statistics, is to square it. This is the well-known proportion of common variance: For correlations of modest size (.20, .30, .40), the square is quite small (.04, .09, .16, respectively), whereas the proportion of nonshared or unique variance ($1 - r^2$) is quite large (.96, .91, .84, respectively). As a consequence, many professionals who are consumers of psychological research conclude at this point that the relationship between the two variables is trivial and that correlations of modest size are of little practical value. For example, after a *Newsweek* story was published involving a National Research Council's report (Wigdor & Garner, 1982) that indicated that cognitive tests were *unbiased* for predicting performance across multiple groups, Bruce Alberts, the president of the National Academy of Sciences, replied in a letter to *Newsweek*, "The prediction for any group is not strong—about nine percent of the variation in job performance." (Alberts, 1994, p. 22).

The perception of triviality also is induced by computing the standard error of estimate, which, in standard score form, is $(1 - r^2)^{1/2}$. This expression is the basis for the commonly quoted statement that the correlation must equal .866 before the error in prediction is reduced by 50% from the error one could make in using a

David Lubinski, Department of Psychology, Iowa State University; Lloyd G. Humphreys, Department of Psychology, University of Illinois, Urbana–Champaign.

Correspondence concerning this article should be addressed to David Lubinski, Department of Psychology, Iowa State University, Ames, Iowa 50010-3180. Electronic mail may be sent via Internet to Lubinski@iastate.edu.

random device. The corresponding value for a correlation of .50 produces only a 13% reduction in errors over chance.

The tendency to dismiss modest correlations is reinforced even further by the inevitable presence of false positive and false negative errors in the prediction of individuals. For example, if both a predictor test and a performance criterion are dichotomized at scores representing minimally satisfactory performance on each measure, the combination of false negative and false positive errors in prediction decreases slowly as the correlation increases from zero to unity. The decrease in false positive and false negative errors is not linear, because the reduction in amount of error positively accelerates with large correlations. *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 1985) require that gains from the reduction of false positive errors must be evaluated against the costs of false negative errors. The harm to persons who would have been "successful" had they not been disqualified by a fallible test is given great weight by critics of tests and by testing experts (Hartigan & Wigdor, 1989).

To summarize, psychological variables generating modest correlations frequently are discounted by those who focus on the magnitude of unaccounted for criterion variance, large standard errors, and frequent false positive and false negative errors in predicting individuals. Dismissal of modest correlations (and the utility of their regressions) by professionals based on this psychometric–statistical reasoning has spread to administrators, journalists, and legislative policy makers. Some examples of this have been compiled by Dawes (1979, 1988) and Linn (1982). They range from squaring a correlation of .345 (i.e., .12) and concluding that for 88% of students, "An SAT score will predict their grade rank no more accurately than a pair of dice" (cf. Linn, 1982, p. 280) to evaluating the differential utility of two correlations .20 and .40 (based on different procedures for selecting graduate students) as "twice of nothing is nothing" (cf. Dawes, 1979, p. 580). It is curious that even among psychologists, all of whom certainly have taken courses in psychological measurement, there are many highly vocal critics who would like to abolish the use of predictor tests. They often assert that the amount of error in prediction is antithetical to democratic values. In truth, what follows may lead one to conclude the opposite.

It should be stressed, however, that we do not argue that concerns about errors in the prediction of individuals are illegitimate. Vocational counselors, clinical psychologists, and others whose roles involve judgments about individuals must be acutely aware of both false positive and false negative errors in predicting individuals. A recommendation that Johnny will become a successful engineer is highly probabilistic. Likewise, looking at the analogue in medicine, the probabilities of false negative and false positive errors in diagnosis and treatment must be known as accurately as possible. A choice between performing a dangerous operation on Jane immediately versus, if the diagnosis is accurate, imminent death requires highly valid diagnostic validities.

In dealing with individuals, giving accurate information to the patient or client is only part of the problem. Helping the person balance the gains and losses from the two kinds of errors also is required. Further, the utility of the decision is not simply an objective function of the probabilities of error. Clients may assess gains and losses differently. And the probability that a civil suit will be brought against practitioners for errors of either type affects the judgment of the practitioner. Errors

made about individuals in clinical settings are inevitable, even when valid information is used by competent consumers of research (cf. Dawes, 1988; Dawes, Faust, & Meehl, 1989; Meehl, 1954, 1956).

Tests are used, however, in ways other than the prediction of individuals or of a specific outcome for Johnny or Jane. And policy decisions based on tests frequently have broader implications for individuals beyond those directly involved in the assessment and selection context (see the discussion later in this article). For example, selection of personnel in education, business, industry, and the military focuses on the criterion performance of groups of applicants whose scores on selection instruments differ. Selection psychologists have long made use of modest predictive correlations when the ratio of applicants to openings becomes large. The relation of utility to size of correlation, relative to the selection ratio and base rate for success (if one ignores the test scores), is incorporated in the well-known Taylor-Russell (1939) tables. These tables are examples of how psychological tests have revealed convincingly economic and societal benefits (Hartigan & Wigdor, 1989), even when a correlation of modest size remains at center stage. For example, given a base rate of 30% for adequate performance and a predictive validity coefficient of .30 within the applicant population, selecting the top 20% on the predictor test will result in 46% of hires ultimately achieving adequate performance (a 16% gain over base rate). To be sure, the prediction for individuals within any group is not strong—about 9% of the variance in job performance. Yet, when training is expensive or time consuming, this can result in huge savings.

For analyses of groups composed of anonymous persons, however, there is a more unequivocal way of illustrating the significance of modest correlations than even the Taylor-Russell tables provide. And when presented in this fashion, dismissal of small correlations becomes almost irrational and allows policy makers to understand more readily their precise utility. Presenting this model is the central message of this article.

Rationale for an Alternative Approach

Applied psychologists discovered decades ago that it is more advantageous to report correlations between a continuous predictor and a dichotomous criterion graphically rather than as a number that varies between zero and one. For example, the correlation (point biserial) of about .40 with the pass–fail pilot training criterion and an ability–stanine predictor looks quite impressive when graphed in the manner of Figure 1a. In contrast, in Figure 1b, a scatter plot of a correlation of .40 between two continuous measures looks at first glance like the pattern of bird shot on a target. It takes close scrutiny to perceive that the pattern in Figure 1b is not quite circular for the small correlation. Figure 1a communicates the information more effectively than Figure 1b.

When the data on the predictive validity of the pilot ability–stanine were presented in the form of Figure 1a (rather than, say, as a scatter plot of a correlation of .40; Figure 1b), general officers in recruitment, training, logistics, and operations immediately grasped the significance of the data for their problems. Because the Army Air Forces were an attractive career choice, there were many more applicants for pilot training than could be accommodated and selection was required. The decision to use the tests that produced results displayed in Figure 1a

Figure 1a

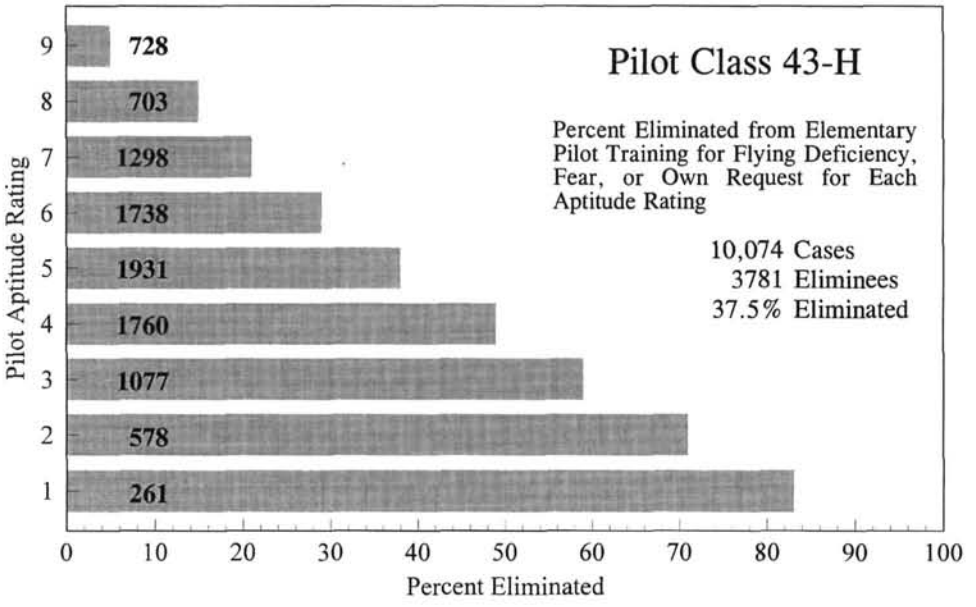

Figure 1b

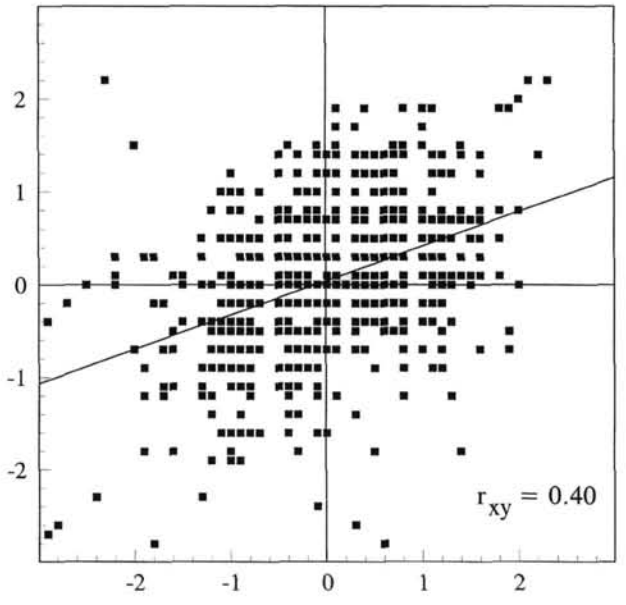

Figure 1. a: Percentage of pilots eliminated from a training class as a function of pilot aptitude rating in stanines. Number of trainees in each stanine is shown on each bar. (From DeBois, 1947). b: A synthetic example of a correlation of .40 (N = 400).

resulted in each of the following (all of which have implications for individuals beyond those that are involved directly in the personnel selection context): (a) fewer training bases (less land, fewer buildings, fewer administrative personnel), (b) fewer instructors (releasing pilots for combat), (c) fewer training aircraft (allowing increased production of combat aircraft), (d) less consumption of aviation gasoline and motor oil (both in short supply), and (e) higher levels of proficiency of the pilots graduating from training. As a corollary, it was also easy to choose the most effective selection device among amount of education, the Army General Classification Test, and the pilot ability–stanine simply by comparing graphs.

Each level of the pilot ability–stanine was, in effect, a mean of all of the scores in a segment of a normal distribution. The proportions passing (or failing) are also means of binary distributions. The graph in Figure 1a represents a "simple" individual differences correlation of modest size, $r_{xy} \cong .40$, but no misinformation was conveyed. That is, the significance of a modest correlation for policy decisions was not hidden by interpretations based on the scatter of individuals in a bivariate plot of two continuous distributions. (A continuous distribution of proficiency underlying the dichotomy of pass–fail is a reasonable assumption, and a good deal of effort was expended in developing such a measure.)

Again, there is an analogue in medicine, particularly in public health medicine. Epidemiological studies are currently being used with confidence as the bases for changes in related policies. Research on smoking is one example, even though the accuracy of predicting death from lung cancer for individuals from the number of cigarettes smoked per day must be very modest indeed.[1] Nevertheless, the accuracy of predicting health effects for groups of persons who are homogeneous with regard to an appropriate measure of amount of smoking is highly predictable. The point is that when individuals are aggregated systematically to form well-defined groups, the harmful effects of smoking are more clearly revealed.[2]

The presence or absence of lung cancer, a dichotomous criterion, limits the figural presentation of a correlation between the predictor and the criterion to a relation between means, just as it does for selection tests and a pass–fail criterion. However, a continuously distributed criterion of degree of emphysema could be correlated with a measure of the amount of smoking in a sample of individuals. Yet, it would be a mistake to present the scatter plot or the correlation as the basis for proposed social action. This is because the presumptive size of the correlation coefficient (predictably small) minimizes the importance of the relation between smoking and health to all but sophisticated users of regression equations. The correlation between means, which is standard in epidemiology, is preferable. We believe it is also preferable in the social sciences when the prediction of groups (defined by homogeneity on a personal or behavioral attribute) is of interest.

---

[1]We are not aware of any published correlations. We have, however, recently secured data from the National Center for Health Statistics (1992) on smoking/nonsmoking during pregnancy and low birth weight. We had to collate information from different tables in this document, but the simple correlation between amount of smoking and low/nonlow birth weight is less than $r_{xy} = .10$.

[2]There is also a less direct analogue in other branches of psychology and in other sciences. Experimental psychologists plot functional relations between several levels of one variable and the means of multiple observations on the dependent variable at each of those points. Biological and physical scientists do the same. For the experimental psychologist, the multiple observations may at times be from a single individual but are more frequently from a series of individuals.

We are not advocating any change (unless it would be to even greater caution) in decisions, predictions, diagnoses, or advice concerning individuals from correlations of modest size, as we have emphasized previously. The utilities associated with false positive and false negative errors about individuals are of prime importance. However, accuracy of group prediction is not affected by the amount of error for individuals. We argue that psychologists who use individual differences data for the prediction of group behavior or status are failing to state their case in a fashion that inspires confidence and action by holding to a methodology aimed at the prediction of individuals. The alternative methodology, graphing and correlating means, allows one to show that psychological tests provide far greater accuracy for predicting social criteria (for purposes of policy formation and change) than is realized by those who see only correlations of modest size, accompanied by large standard errors of estimation.

Correlating Bivariate Means

Our proposal is this: Present all correlations that may serve as a basis for social and policy decisions or theoretical analyses of group trends as graphs of the regression of criterion means on predictor means. These bivariate means may be derived by parsing the predictor into equal intervals and computing predictor–criterion means within each interval. In order to interpret the data accurately, as well as to use them in new samples from the same population, the graphs should be accompanied by the regression equation that makes use of the statistics based on means:

$$\hat{\bar{y}}_i = [r_{\bar{x}\bar{y}}(s_{\bar{y}}/s_{\bar{x}})\bar{x}_i] + [\bar{\bar{y}} - r_{\bar{x}\bar{y}}(s_{\bar{y}}/s_{\bar{x}})\bar{\bar{x}}] \quad (1)$$

where

$r_{\bar{x}\bar{y}}$ = correlation between bivariate means,
$\bar{x}_i$ = successive group means on the predictor,
$\hat{\bar{y}}_i$ = predicted group mean on the criterion at each $\bar{x}_i$,
$s_{\bar{y}}$ = standard deviation of y means,
$s_{\bar{x}}$ = standard deviation of x means,
$\bar{\bar{y}}$ = mean of the y means,
$\bar{\bar{x}}$ = mean of the x means.

In what follows, we illustrate the extent to which regression equations based on predictor–criterion group means (derived by parsing the predictor into equal intervals, computing predictor and criterion means within each interval, and then correlating these bivariate distributions of means) can provide nearly error-free prediction of the group performance of new samples from the same population. The essential conditions for this degree of accuracy are linearity of regression in the population and a sample of sufficient size to produce stable estimates of its slope. Given these conditions, the correlation required by Equation 1 inevitably approaches 1.00, even though the slope of the line obtained from that equation reproduces the slope of the regression equation for individuals. Although these illustrations provide no information that is not implicit in regression equations that use correlations computed on samples of individuals (unless the regression is

curvilinear; see the material that follows), we believe that they highlight more effectively the value of psychological predictors for forecasting trends and making policy decisions about groups of people. In effect, we are advocating an epidemiological approach to the presentation of psychological predictive data when the objective is similar to developing policy in public health.

Some Empirical Examples

The data. We analyzed data from the Project Talent Data Bank (Wise, McLaughlin, & Steel, 1979), which contains almost all the students in a stratified random sample of the nation's high schools in 1960. It contains four cohorts, Grades 9 through 12, with approximately 100,000 per cohort, the individuals of which were administered several dozen conventional status and individual differences measures. We will illustrate our points using data from the 12th-grade cohort. The variables chosen for analysis were Talent's (Flanagan et al., 1962) Intelligence Composite, the predictor, and three criterion measures: general information, socioeconomic status, and high school mathematics.[3]

Defining screening and calibration samples for analysis. Following the screening and calibration nomenclature of Lord and Novick (1968, p. 285), we divided the 12th-grade cohort by gender into two screening samples (used to compute the regression equations) and two calibration samples (used to cross-validate these equations). In the male–female screening samples, means and standard deviations were computed for all predictor and criterion variables (see Table 1). For each gender, screening distributions of the Intelligence Composite were systematically parsed into equal intervals, approximately .20 standard deviation units each, extending from the mean in both directions. (This parsing was discontinued when the number of individuals fell below 2% of the sample in each of the two extreme class intervals.[4]) The modifier *approximately* is used because some rounding error was involved (each interval extended 10 raw-score

---

[3]*The predictor–criterion measures*: Project Talent's Intelligence Composite contains three components: reading comprehension, arithmetic reasoning, and abstract reasoning (283 possible point range). The Intelligence Composite comes closest to matching the content found on conventional measures of general intelligence—the Stanford-Binet Scale (Terman & Merrill, 1960) and the various Wechsler (1974) tests of general intelligence. This composite will serve as the predictor variable for all of our analyses. The following three variables will be used as criteria. *General information* (143 possible points) consisted of a broad range of information tests, not particularly linked to high school experiences: art, law, medicine/health, engineering, architecture, journalism, foreign travel, military, accounting/business/sales, practical knowledge, clerical, the Bible, colors, etiquette, hunting, fishing, outdoor activities, photography, sedentary games, theater and ballet, foods, and general vocabulary. *Socioeconomic status* (135 possible points) was a conventional measure of socioeconomic status, which included value of home, family income, number of books in the home, appliances, father's occupation, parent's education. *High school mathematics*: (24 possible point range) this is a measure of math achievement of all kinds of mathematics generally taught up to and including 9th grade; the primary emphasis of this test is on algebra, but other topics include fractions, percents, decimals, square roots, and elementary measurement formulas.

[4]In addition to the problems of relatively small sample sizes in the tails of the distribution of general intelligence, the almost perfect linear regressions of our dependent measures on the predictor in the main body of the distribution actually reverse their direction in the lower tail. The small subset of persons responsible for the reversal are the subject of detailed analyses (Humphreys, Lubinski, & Yao, 1993).

Table 1
Raw Score Means and Correlations for Predictor and Criterion Variables by Gender for the Screening and Calibration Samples

Measure	Intelligence composite	SES	General information	High school math	Men $\bar{x}$	Men *SD*	Women $\bar{x}$	Women *SD*
			Screening sample					
Intelligence composite	—	.40	.78	.65	185.87	52.94	180.96	51.01
SES	.42	—	.45	.36	99.09	10.05	98.83	9.76
General information	.78	.46	—	.57	77.33	20.41	73.73	18.59
High school math	.72	.39	.61	—	12.33	5.66	10.28	4.85
			Calibration sample					
Intelligence composite	—	.40	.79	.66	186.24	53.23	180.54	51.32
SES	.42	—	.45	.36	99.12	10.03	98.87	9.72
General information	.77	.45	—	.58	77.41	20.58	73.71	18.72
High school math	.73	.40	.62	—	12.40	5.64	10.28	4.89

Note. Correlations for female research participants are above the diagonal; correlations for male participants are below. Screening sample—males: n = 17,358, females: n = 18,337; Calibration sample—males: n = 17,255, females: n = 18,512. SES = socioeconomic status.

units for both genders, and this only approximates .20 standard deviations): For the boys, $\bar{x}$ = 185.87, SD_x = 52.94, and the predictor was parsed at and above 186 and at and below 185; for the girls $\bar{x}$ = 180.96, SD_x = 51.01, and the predictor was parsed at and above 181 and at and below 180.

The selection of 20 intervals on the predictor, as opposed to many fewer intervals needed to illustrate the methodology, was based on our desire to illustrate the fit of linear regression with considerable precision. The size of our samples and the wide range of scores in our quasi-continuous distributions allowed the use of many intervals without any appreciable reduction in the correlation between means.

These same raw-score intervals were applied to the gender-equivalent calibration samples; bivariate means were computed for the calibration samples in the same manner. Because our measurement instruments were scaled in arbitrary units, we wanted to view our regression lines in terms of more familiar units. Thus, all predictor and criterion means, within each interval and across both screening and calibration samples, were transformed into approximations of standard scores using the means and standard deviations of the gender-equivalent screening samples. Finally, because of the amount of data we are about to report, our results and discussion will focus on scatter plots of the calibration samples and cross-validation coefficients comparing their regressions to those of the screening samples.

Results and Discussion

Using data from the male and female screening samples, six regression equations were computed 2 (gender) × 3 (criteria); they determined the regression lines found on Figures 2A, 2B, and 2C (for the girls) and Figures 2D, 2E, and 2F

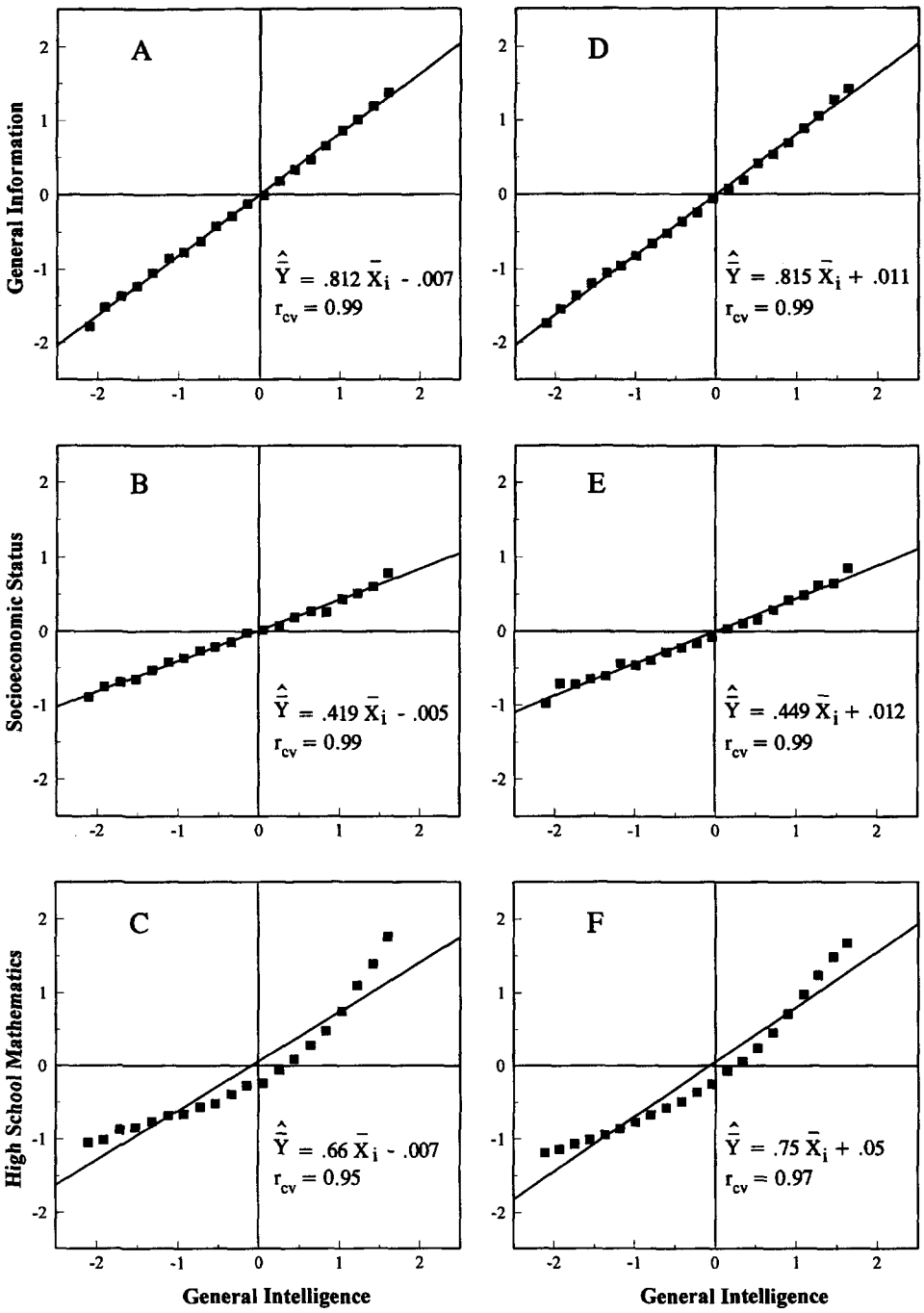

Figure 2. Regressions of means of general information, socioeconomic status, and high school math on the means of general intelligence. Data for the female students are in the left-hand panels; data for the male students are in the right-hand panels. Cross-validation coefficients r_{cv} were computed by correlating the calibration samples observed bivariate means with the predicted values.

(for the boys). Each figure, in addition to the regression line ($\hat{\bar{y}}_i$), contains the bivariate means (the observed $\bar{y}$s), for the calibration samples. Finally, cross-validation (r_{cv}) coefficients also are found in the lower-right quadrant of each figure (computed by correlating the predicted values, found on the regression line, with the calibration samples' observed bivariate means). All of these values are just under unity. All correlations were computed using unit weight for each pair of means.

Interpreting these regressions. The analyses just discussed reveal that group performance and status can be predicted with remarkable precision even when "simple" predictor–criterion correlations are moderate in size. As the figures illustrate, group performance on general information (Figures 2A and 2D) and group level on socioeconomic status (Figures 2B and 2E), estimated from group means on the intelligence composite can be predicted with near certainty. The cross-validation equations required for predicting in new samples contain $r_{cv} >$ .990 for general information and socioeconomic status. The slope of the bivariate values, for all four calibration samples, is approximately equal to the slope of the regression line, $\hat{\bar{y}}_i$, computed on the screening samples. Because the units of measurement were approximately standard, the figures also reveal the magnitude of the predictor–criterion relationship. The slope of the line for general information is nearly twice as steep as the slope for socioeconomic status, which of course would be expected from the simple correlations (r_{xy}) found in Table 1. The slope of the regression line for the grouped data, namely, $b = r_{\bar{x}\bar{y}} s_{\bar{y}}/s_{\bar{x}}$, should equal to a close approximation the simple correlation r_{xy} of conventional regression analysis (see Table 1). Higher accuracy can be obtained by weighting pairs of means by the size of the defined group. Also, to the extent that $r_{\bar{x}\bar{y}} \rightarrow 1.0$, one can estimate the simple correlation using the ratio of the standard deviations of the means: $s_{\bar{y}}/s_{\bar{x}}$.

That the simple correlation r_{xy} approximately equals $s_{\bar{y}}/s_{\bar{x}}$. when $\bar{x}$ and $\bar{y}$ are standardized is important, because these values estimate the difference in group means on the criterion, in standard deviation units, for every standard deviation difference between two groups on the predictor. So, for modest predictor–criterion correlations of, say, .20 or .40, every standard deviation difference between predictor group means corresponds to a precise .20 or .40 standard deviation difference, respectively, between the criterion group means. A group standard-score difference of .20 or .40 on a substantively significant criterion variable (e.g., law school bar examinations or National Board Examinations in medicine) can have great utility (Cronbach & Gleser, 1965; Hunter & Schmidt, 1983; Schmidt, Ones, & Hunter, 1992; Taylor & Russell, 1939).[5]

[5]The graph of the regression of criterion means on predictor means can also provide a contrast between the objectives of predicting performance of individuals and of groups. The two objectives do not have to be considered in every application, but they should certainly not be confused. For individuals, the conventional *standard error of estimate* [i.e., $S_y(1 - r_{xy}^2)^{1/2}$] can be used to place a confidence interval about the regression of criterion means on predictor means. As sample size increases, the confidence interval becomes more stable, but the accuracy of predicting an individual's criterion status is not affected. For groups, the size of the *sampling error* about the regression line determines the size of an alternative confidence interval about the same regression line and can be approximated by the following formula: $S_y(1 - r_{xy}^2)^{1/2} \div N_i^{1/2}$ (where N_i = the sample size of the ith group defined on the predictor). This confidence interval is always smaller than the one required for predicting the performance of an individual, and its size decreases monotonically as a function of N. Here, as sample sizes increase, the amount of gain in mean criterion performance accompanying a group's unit gain on the predictor mean can be determined with almost perfect accuracy.

Revealing nonlinear trends. Figures 2C and 2F illustrate how this methodology may be used to uncover curvilinear trends that might go undetected by conventional regression analyses. The simple correlations between general intelligence and high school mathematics (.72 for boys and .65 for girls) would generate scatter plots with regressions not readily distinguished from linearity. The cross-validated correlations between means are .97 and .95 for boys and girls, respectively! Yet these regressions are clearly curvilinear and became obviously so in the scatter plot of bivariate means. Accuracy in prediction was only attenuated a little using the linear model (down from .99 to .95 for the girls), but clearly a rational, curvilinear function could be fitted to these data to enhance prediction.

Other approximations to this approach. Our demonstration of stability in cross-validation analyses is more elaborate than the requirements for broad use of correlations between means. The initial computation of the correlations in individual differences data (i.e., the simple correlations found in Table 1) are not required. It is important, however, to have N sufficiently large and the number of means sufficiently small to minimize sampling error. The regression equation needed for predicting individual performance is not a prerequisite for predicting mean criterion performance or mean criterion status. A continuous predictor can be converted into N class intervals and a mean criterion score computed for each interval. The product–moment correlation between the two distributions of means then can be computed and used in the regression equation appropriate for means. The regression of means on means cannot be linear unless the regression of criterion scores on predictor scores is also linear. For example, when well-designed ability tests are used to predict proficiency criteria with desirable measurement properties (Carroll & Horn, 1981; Keating & Stanley, 1972; Linn, 1982; Schmidt & Hunter, 1992), there are few regressions in samples typical of those psychologists ordinarily use in which the hypothesis of linearity can be rejected.

Moreover, when the regression in the individual differences data is not presented (e.g., the simple r_{xy}s), as in Figure 1a, nothing is actually hidden. If both the predictor and criterion have been converted to distributions having means of zero and standard deviations of 1.0, the size of the individual differences correlation can be estimated directly from the graph of the functional relation or the slope of the regression line. It also can be estimated with nugatory error from the ratio of standard deviations of the distributions of means: $s_{\bar{y}}/s_{\bar{x}}$.

When raw scores are in meaningful units, we recommend that the regression equation for the means be in terms of the raw score units. In this instance, additional information is required in order to estimate the simple r_{xy} correlation, but knowledge of the correlation is not essential. If experts agree that the relation portrayed has utility, it does not matter whether the original slope included a correlation of .20 or .80. Nonetheless, when the regression of the criterion on the predictor is linear, if one has available the standard deviation in the individual differences sample of the criterion scores, the ratio of the standard deviation of the criterion means to the former standard deviation provides an estimate of the simple r_{xy} correlation that is only slightly inflated in large samples.

Finally, although we have chosen to illustrate our proposed methodology using systematically parsed intervals on the predictor variable to define our groups, groups can be formed in multiple ways. Some more naturally occurring groups (for

which our methodology is also applicable) are college majors, occupation, political affiliation, psychiatric types, and race. Bivariate predictor–criterion means for these groups (based on mean predictor status and mean criterion status) are easily accommodated by the proposed regression of criterion means onto predictor means.

Sampling error. We bypassed earlier the problem of estimating sampling error around our regression lines. In samples of almost 25,000 cases, sampling stability is a trivial problem for this methodology. However, we recommend the use of this technique in much smaller samples. Given homoscedasticity in the y scores in the arrays defined by x, the sampling error of the slope of the regression line in individual data is given by the following:

$$SE_{\text{slope}} = S_y(1 - r_{xy}^2)^{1/2}/S_x(N - 2)^{1/2}. \quad (2)$$

Because the slope of the regression using bivariate means is essentially identical with the slope in individual data, this standard error can be used for the slope of the regression of means on means. The simple regression of y on x has its origin in $\bar{x}$, whose sampling error is well known. Because $\bar{x}$ is also nearly identical to the mean of the means, $\bar{\bar{x}}$, as just computed, it also can be used for the mean of the means (an exact estimate would be obtained by computing a weighted mean). Given that the total sample size appears in both standard errors, samples do not need to be nearly as large as ours in order to have confidence in the regression of means on means in estimating group outcomes in new samples. Indeed, if N increased to merely several hundred, as it often does for the prediction of clinical outcomes (Dawes et al., 1989; Meehl, 1954, 1986), college grades (Jensen, 1980; Linn, 1982; Stanley, 1971), military proficiency criteria (DuBois, 1947; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990), or validity generalization studies (Hunter, 1980; Hunter, Schmidt, & Hunter, 1979; Schmidt & Hunter, 1992), the amount of error in prediction would be modest.

Magnitude appraisals in experimental psychology. To more fully appreciate the magnitude of small correlations, it is instructive to examine a common practice among experimental psychologists. Following Cohen (1988), experimenters routinely refer to *effect size differences* (standard deviation) between control and experimental groups as: small $\geq .20$, medium $\geq .50$, and large $\geq .80$. Cohen (1988, p. 22) provided a formula for converting effect size differences (or *ds*) into correlational units[6]: $r = d \div (d^2 + 4)^{1/2}$. But the significance of this transformation for evaluating the magnitude of predictor–criterion covariation in individual differences research is underappreciated. Small, medium, and large effect size differences translate into correlations of .10, .24, and .37, respectively! Given *ds* of .80 are considered large when uncovered by *experimental* psychologists (examining differences between experimental and control groups), we recommend giving such values commensurate attention when uncovered by *differential* psychologists (examining differences between groups as a function of predictor–criterion covariation). For further treatment of this idea and how it pertains to this article, readers are referred to Lubinski and Humphreys (in press).

[6]The more general formula for the transformation exchanges "$1/pq$" with 4, where p = proportion of Group 1 in combined Group 1 and Group 2 populations and $q = 1 - p$ (i.e., proportion of Group 2 in combined Group 1 and Group 2 populations). Thus: when $p = q$, $1/pq = 4$.

Conclusion

If a *nonzero* correlation computed across individuals has an approximately linear slope in the population, the correlation between bivariate means ($r_{\overline{xy}}$), that is, means of the criterion variable and means of the segments of the predictor in which the former are obtained, will approach unity as sample size increases—and will do so quickly. Given that group performance can be predicted with this degree of accuracy, the utility of the predictor (for either theoretical or applied concerns involving groups) can be grasped more readily in terms of the amount of gain on the criterion that accompanies a *unit of gain* on the predictor. (If the regression in the population is nonlinear but monotonic, a linear function may still provide a good fit, but a rational function can be substituted to achieve higher accuracy. If raw scores are not in meaningful units, investigators can normalize one or both distributions and convert nonlinearity to approximate linearity in many cases.)

A small gain on a criterion for a unit of gain on the predictor, as long as it is predicted with near-perfect accuracy, can have high utility. Also, given the assumption of approximate linearity in the population of individual persons, *there are no false negative or false positive groups.* Given a correlation between means near unity, the standard error of estimation for the prediction of groups is almost zero. Viewers, in looking at graphic presentations of data involving *grouped* bivariate means, are not distracted or misled, as they often are in conventional presentations of individual differences data, by the amazing scatter of scores about regression lines.

The precision that can be achieved with individual differences variables for predicting group trends versus individual performance or status is often underappreciated by psychologists, policy makers, and other consumers of psychological research. Greater appreciation is attained when individual data points are systematically grouped to form bivariate means, to replace the constituent trees with a view of the forest.

References

Alberts, B. (1994, 21 November). Letters. *Newsweek,* p. 22.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Carroll, J. B., & Horn, L. L. (1981). On the scientific bases of ability testing. *American Psychologist, 36,* 1012–1020.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

DuBois, P. H. (1947). *The classification program* (Report No. 2 of the Research Reports from the Army Air Forces Aviation Psychology Program). Washington, DC: U.S. Government Printing Office.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldbert, I. (1962). *Design for a study for American youth.* Boston: Houghton Miffin.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Correlates of some curious regressions on a measure of general intelligence. *Journal of School Psychology, 31,* 385–405.

Hunter, J. E. (1980). *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.

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

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

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Keating, D. P., & Stanley, J. C. (1972). Extreme measures for the exceptionally gifted in mathematics and science. *Educational Researcher, 1,* 3–7.

Linn, R. L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences and controversies* (pp. 335–388). Washington, DC: National Academy Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley

Lubinski, D., & Humphreys, L. G. (in press). Incorporating general intelligence into epidemiology and the social sciences. *Intelligence.*

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. A. (1990). Project A validation results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335–353.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1956). Wanted—a good cookbook. *American Psychologist, 11,* 263–272.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370–375.

National Center for Health Statistics. (1992). Advance report of new data from the 1989 birth certificate. *Monthly vital statistics report, 40* (Suppl. 12), pp. 12–16. Hyattsville, MD: Public Health Service.

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

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual Review of Psychology, 43,* 627–670.

Stanley, J. C. (1971). Predicting college success of the educationally disadvantaged. *Science, 171,* 640–647.

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

Terman, L. M., & Merrill, M. A. (1960). *Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M.* Boston: Houghton Mifflin.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for children.* New York: Psychological Corporation.

Wigdor, A. K., & Garner, W. R. (1982). *Ability testing: Uses, consequencies, and controversies.* Washington, DC: National Academy Press.

Wise, L. L., McLaughlin, D. H., & Steel, L. (1979). *The Project Talent data handbook.* Palo Alto, CA.: American Institutes for Research.