Restriction of Range Corrections When Both Distribution and Selection Assumptions Are Violated

Alan L. Gross and Lynn Fleischman
City University of New York

In validating a selection test (x) as a predictor of y, an incomplete xy data set must often be dealt with. A well-known correction formula is available for estimating the xy correlation in some total group using the xy data of the selected cases and x data of the unselected cases. The formula yields the r_{yx} correlation (1) when the regression of y on x is linear and homoscedastic and (2) when selection can be assumed to be based on x alone. Although previous research has considered the accuracy of the correction formula when either Condition 1 *or* 2 is violated, no studies have considered the most realistic case where both Conditions 1 and 2 are simultaneously violated. In the present study six real data sets and five simulated selection models were used to investigate the accuracy of the correction formula when neither assumption is satisfied. Each of the data sets violated the linearity and/or homogeneity assumptions. Further, the selection models represent cases where selection is not a function of x alone. The results support two basic conclusions. First, the correction formula is not robust to violations in Conditions 1 and 2. Reasonably small errors occur only for very modest degrees of selection. Secondly, although biased, the correction formula can be less biased than the uncorrected correlation for certain distribution forms. However, for other distribution forms, the corrected correlation can be less accurate than the uncorrected correlation. A description of this latter type of distribution form is given.

In validating a selection test (x) as a predictor of a criterion variable (y) there is typically a missing data problem; although x scores are available for all examinees, y is observed only for the selected cases. Thus, the xy relationship cannot be directly inferred for all cases. The problem has been referred to in the literature as the ''restriction of range problem.'' For a specific example, consider the problem of validating the Law School Aptitude Test (x) as a predictor of first-year law school grades (y). Paired xy data are available for selected applicants, but only x scores are observed for rejected applicants. Although the xy relationship can readily be described for the selected cases, interest is primarily in inferring the xy relationship for the total set of applicants. Thus, the statistical problem is to infer the unknown validity of x for the applicant group, given the xy relationship observed in the selected group.

A widely known ''correction formula'' (Lord & Novick, 1968, p.143) is available for estimating the applicant group xy correlation (r_{yx}) given the xy correlation observed in the selected group (r_{yxs}) and the ratio of the variance of x in the selected and applicant groups (s^2_{xs}/s^2_x). The estimate is given as

$$r_{yx} = r_{yxs} \,/\, [r^2_{yxs} + (s^2_{xs}/s^2_x)(1 - r^2_{yxs})]^{\frac{1}{2}} \qquad [1]$$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 2, Spring 1983, pp. 227–237
© Copyright 1983 Applied Psychological Measurement Inc.
0146-6216/83/020227-11$1.55

When selection is based at least in part upon x, the ratio (s^2_{xs}/s^2_x) will be less than unity. Thus, in terms of Equation 1, the xy correlation computed in the selected group (r_{yxs}) will typically underestimate the total group correlation (r_{yx}).

An important question for any practitioner to ask is under what conditions will Equation 1 yield "accurate" values for r_{yx}? A simple condition (Gross, 1982) can be described in the following manner: Suppose the best fitting linear regression line were to be considered for predicting y from x in the applicant group. Let b and s_e represent the slope coefficient and residual standard deviation, respectively. Further, suppose the best fitting linear regression line were to be considered for predicting y from x in the selected group. Let b_s and s_{es} represent the slope coefficient and residual standard deviation, respectively. If the ratio W equals unity, where

$$W = (s_e/s_{es})/(b/b_s) \qquad [2]$$

Equation 1 will yield the exact value for the applicant group correlation (r_{yx}). For $W < 1$ Equation 1 will yield an underestimate of r_{yx}, and for $W > 1$ it will yield an overestimate. In situations where the form of the regression of y on x in the applicant group is exactly linear and homoscedastic (Condition 1), and selection is solely a function of x (Condition 2), W will equal unity and Equation 1 will be exact. More specifically, given Conditions 1 and 2, it follows that b and b_s will be equal, s_e and s_{es} will be equal; consequently, W will be equal to 1. However, it is theoretically possible for Equation 1 to yield accurate values when Conditions 1 and 2 do not hold. For example, situations may occur where the regression of y on x may be nonlinear and heteroscedastic, and selection may be a function of an entire set of variables rather than x alone. In such cases, s_e and s_{es} will in general differ, as will b and b_s. However, differences between s_e and s_{es} can be offset by corresponding differences between b and b_s, with the end result that W can be close to 1. It is important to note, however, that when neither Conditions 1 nor 2 holds, it may be difficult to predict the value of W.

In investigating the accuracy of the correction formula given by Equation 1, researchers have studied cases where either Condition 1 or Condition 2 does not hold. The bulk of the research has considered violations in Condition 1 (e.g., Greener & Osburn, 1979, 1980; Novick & Thayer, 1969.) Basically, these latter studies have shown that the formula yields reasonably accurate values for moderate degrees of selection on x but becomes increasingly less accurate with increasing levels of selection. Further, depending upon the manner in which Condition 1 is violated, Equation 1 can either underestimate or overestimate r_{yx}. Only one study has considered violations in Condition 2. Linn (1968) considered the case where selection is based on some third variable rather than x. However, the relationships among all variables were assumed to be linear and homoscedastic. The results suggested that small degrees of underestimation (approximately of the order of 10%) occur when the correction formula is used to represent r_{yx}.

It is important to note that none of these previous studies has considered the accuracy of Equation 1 when both Conditions 1 and 2 are simultaneously violated. In many real selection problems, it is most realistic to assume that neither condition will be satisfied. For example, it has been suggested (Lord & Novick, 1968) that it is not uncommon to find that the regression of y on x flattens out for extreme x values, and the variance of y given x decreases for extreme x values. Simultaneously, in many problems the selection of applicants is not a function of a single x variable alone. For example, law school admissions would typically be based on LSAT scores and additional variables such as college grades, letters of recommendation, and performance during a personal interview. Thus, it is of practical value to consider the accuracy of Equation 1 when *both* the distribution and selection assumptions given by Conditions 1 and 2 are simultaneously violated.

The present paper is directed at this issue. The method consists of examining real data sets where the linearity and homoscedasticity assumptions are clearly violated. Five different selection models are then applied to each data set producing various selected groups. The accuracy of the correction formula is evaluated by comparing the actual correlation to the value computed in terms of Equation 1. More specifically, the following questions are considered: (1) Is the corrected correlation robust to simultaneous violations in the distribution assumptions (linearity and homoscedasticity) and selection assumptions (selection based on x alone)? (2) Given that the distribution assumptions are violated, what is the additional effect of not selecting on x alone? (3) Does the correction formula yield a more accurate estimate than the uncorrected correlation (r_{yxs}) computed in the selected group?

Method

Data Sets

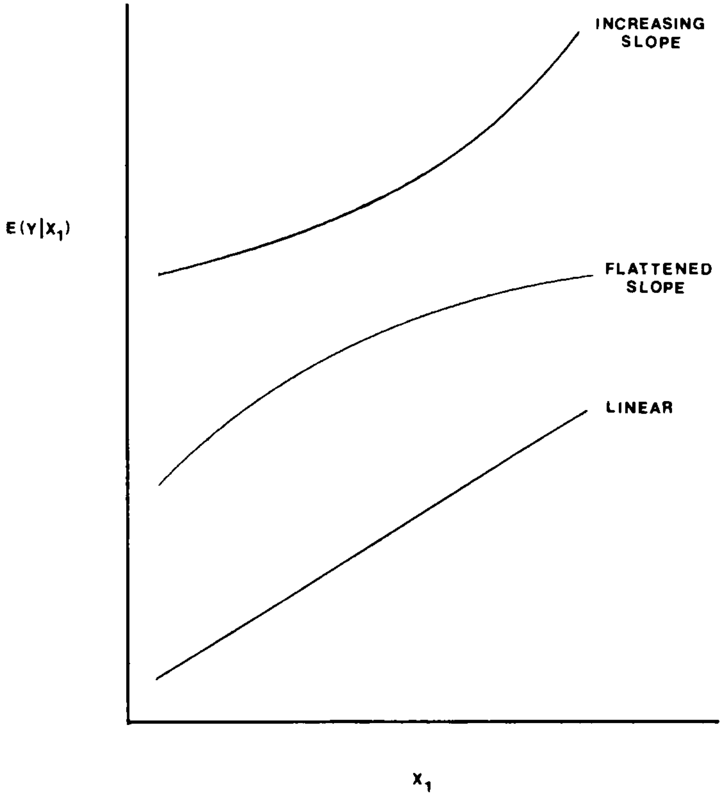
Test score and biographical data were available on a sample of N = 913 school children. The variables consisted of scores on standardized achievement tests, aptitude tests, as well as age and grade level. The achievement tests consisted of the 15 scales of the Comprehensive Tests of Basic Skills (CTB/McGraw Hill, 1977). Each of these tests was represented in terms of grade-equivalent scores. The aptitude tests consisted of the verbal and nonverbal scales of the Short Form Test of Academic Aptitude (Sullivan, Clark, & Tiegs, 1970). The latter two measures were in standardized form. Six trivariate data sets consisting of y, x_1, and x_2 were constructed.

In each of the six data sets the assumptions of linearity and/or homoscedasticity for the regression of y on x_1 was violated. In Data Set 1 the regression was linear but not homoscedastic. A plot of the conditional variances of y as a function of x_1 showed a bimodal form. A schematic representation of this ''bimodal'' form can be seen in Figure 2. In Data Set 2 the regression was both nonlinear and heteroscedastic. The nonlinearity was due to a flattening of the regression slope for higher values of x_1. Figure 1 depicts this ''flattened slope.'' The conditional variances in Data Set 2 were a decreasing function of x_1. This ''decreasing'' form is depicted in Figure 2. For Data Set 3 the regression was both nonlinear and heteroscedastic. The slope of the regression curve increases for higher x_1 values. The slope of this curve (''increasing slope'') is depicted schematically in Figure 1. The plot of the conditional variances of y as a function of x_1 exhibited the ''decreasing'' form shown in Figure 2. In Data Set 4 the regression was linear but not homoscedastic. The plot of the conditional variances exhibited the ''unimodal'' form depicted in Figure 2. Basically, the conditional variance was smallest for extreme x_1 values and greatest for middle range values. For Data Set 5 the regression was nonlinear (flattened slope) and the conditional variances were heteroscedastic (decreasing). Finally, for Data Set 6 the regression was nonlinear (increasing slope) and the conditional variances were heteroscedastic (unimodal).

For each data set the correlation between x_1 and x_2 was approximately .60. Further, in all six data sets the r_{yx1} and r_{yx2} correlations were approximately equal. For Data Sets 1, 2, and 3, the r_{yx1} and r_{yx2} correlations were low, ranging from 36 to .42. For Data Sets 4, 5, and 6 the correlations were moderate, ranging from .56 to .63.

In addition to describing the six data sets in terms of the form of the regression of y on x_1, it is also of interest to consider the form of the regression of y on x_2 and x_1 on x_2. An analysis of the six yx_2 and the six x_1x_2 regressions showed that nonlinear and heteroscedastic forms predominated. Linear regressions were observed only for the yx_2 regressions in Data Sets 1 and 5. Further, 10 of 12 regressions were heteroscedastic, the x_1x_2 regression in Data Sets 1 and 2 being the only exceptions. Finally, it should be noted that in no case was the regression both linear and homoscedastic.

Figure 1
Schematic Representation of the Shape of the Regression Curves



In all six data sets, the lack of linearity was investigated by fitting increasingly higher order polynomial regression functions to the data. The occurrence of significant higher order terms indicated the presence of nonlinearity. Heteroscedasticity was empirically investigated by grouping the independent variable into a set of ordered intervals and then graphing the conditional variance of the dependent variable for each interval.

The Selection Process

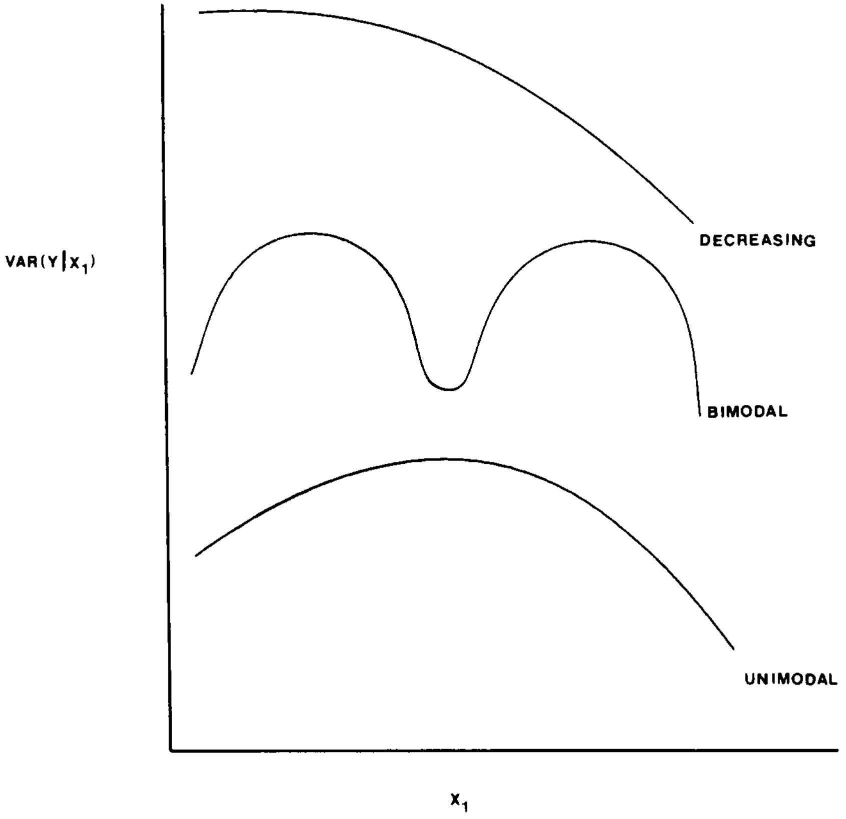
Five different selection processes were simulated using each of the six data sets. Each of these models was based on the selection of cases in terms of x_1 and/or x_2. The general procedure can be described in the following manner: Let R, R_{x_i}, S_{x_i} represent, respectively, the events of being rejected (i.e., not selected), being rejected in terms of x_i, and being selected in terms of x_i, $i = 1,2$. The overall probability of being rejected (not selected) can then be represented as a function of x_1 and x_2 as follows:

$$P(R) = P(R_{x1}) + [1 - P(R_{x1})][P(R_{x2}|S_{x1})] \qquad [3]$$

Thus, the rejection process can be viewed as a type of additive function of x_1 and x_2. The relative importance of x_1 and x_2 in the selection process can be assessed by considering the following ratio:

Figure 2

Schematic Representation of the Conditional Variances of y as a Function of x_1



$$k = P(R_{x1})/([1 - P(R_{x1})][P(R_{x2}|S_{x1})]) \quad [4]$$

As k approaches zero, the selection process becomes solely a function of x_2. As k becomes large, selection becomes solely a function of x_1. Finally, for intermediate k values, selection is based to varying degrees on both x_1 and x_2. To understand this statement, let the probability of being rejected be expressed in terms of x_1 as a function of k. Using Equations 3 and 4 gives

$$P(R_{x1}) = [k/(k + 1)] \, P(R) \quad [5]$$

Thus, for example, for very large k, the ratio $k/(k + 1)$ will be approximately 1, and $P(R_{x1})$ and $P(R)$ will be approximately equal. Thus, rejection can be viewed as a function of x_1 alone. On the other hand, as k approaches zero, the ratio $k/(k + 1)$ will approach zero and rejection will not be a function of x_1, but rather a function of x_2.

The value of k was chosen so that the ratio $k/(k + 1)$ assumed the following values: 0, .25, .50, .75, 1. In addition, the overall proportion selected, $P(S) = 1 - P(R)$ was varied from .20 to .90. For $k/(k + 1) = 0$, selection was performed by choosing the highest scoring cases on x_2. For $k/(k + 1) = 1$, selection was performed by selecting the highest scoring cases on x_1. For $k/(k + 1) = .25, .50, .75$, a multiple cutoff procedure based on both x_1 and x_2 was employed. The selection procedure consisted of selecting all cases for which $x_1 > c_1$, $x_2 > c_2$, where c_1 and c_2 were chosen so that the desired value of $k/(k + 1)$ was achieved. For example, for $P(R) = .60$, $k/(k + 1) = .25$, the value of $P(R_{x1})$ was computed from Equation 5 to be (.25)(.60) = .15. The data set was then rank ordered on x_1 and the highest scoring $1 - P(R_{x1}) = .85$ of the $N = 913$ cases were considered to be selected. The number selected on x_1 was thus $NS_{x1} = 776$. A given number (NS_{x2}) of these cases was selected in terms of x_2. Using Equation 4 and noting that $k = .33$ for $k/(k + 1) = .25$, the probability of being rejected in terms of x_2 after having been selected on x_1, $P(R_{x2}|S_{x1})$, was computed to be .15/[(.33) (.85)] = .53. Thus, the proportion $1 - P(R_{x2}|S_{x1}) = .47$ of the cases scoring highest on x_2 were considered to be the final selected group of size (.47) (776) = 365. As the ratio $k/(k + 1)$ was increased to the values of .50 and .75, selection became increasingly a function of x_1. For example, for $k/(k + 1) = .75$, 45% of the cases were rejected in terms of x_1, and only 27% of these were then rejected in terms of x_2. Finally, for $k/(k + 1) = 1$, the highest scoring cases on x_1 alone were selected.

In summary, for a given proportion selected, $P(S) = .20, .30, \ldots, 90$, five different selection procedures were applied to a data set. The cases where $k/(k + 1) = 0, .25, .50, .75$ represent violations in the assumption that selection is based on x_1 alone. Further, the lower the value of $k/(k + 1)$, the greater the degree of violation in this assumption. The case where $k/(k + 1) = 1$ provided a selection model that is based on x_1 alone and provided for a type of control condition.

Data Analysis

For each of the six data sets, five selection models and eight degrees of selection $P(S) = .20, .30, \ldots, 90$, the accuracy of the correction formula was assessed in terms of a measure of percentage error. More specifically, given a selected group, the uncorrected correlation (r_{yx1s}) and variance of x_1 (s^2_{x1s}) were computed. Using the correction formula and the variance of x_1 in the total group (s^2_{x1}), the total sample correlation was estimated. Denoting the estimate of the total sample correlation given by the correction formula as $\hat{r}_{yx1}$ and the true value as r_{yx1}, the percentage error for the correction formula was computed as follows:

$$P_C = (\hat{r}_{yx1} - r_{yx1})/r_{yx1} \qquad [6]$$

In addition, the percentage error for using the uncorrected correlation (r_{yx1s}) to estimate r_{yx1} was computed as follows:

$$P_U = (\hat{r}_{yx1s} - r_{yx1})/r_{yx1} \qquad [7]$$

Results

In Table 1 the P_C and P_U values are reported for each of the six data sets. For each data set the percentage errors are presented as a function of the proportion selected ($P(S) = .20, .30, \ldots, 90$), and the value for $k/(k + 1) = 0, .25, .50, .75, 1$. Negative percentage errors denote underestimation of the total sample correlation (r_{yx1}), whereas positive values denote overestimation.

An inspection of Table 1 shows that, in general, the correction formula cannot be considered to be robust with respect to simultaneous violations in the underlying distribution and selection assumptions. If the cases are considered where $k/(k + 1) \leq .75$, i.e., the cases where selection is not based on x_1 alone, reasonably small percentage errors are observed only for mild degrees of selection. More specifically, for 90% selected $|P_C| \leq .14$, and for 80% selected $|P_C| \leq .21$. However, as the proportion selected decreases, the accuracy of the formula deteriorates. For example, even when 70% are selected, percentage errors as large as 37% are observed.

Table 1
Percentage Error for Corrected (C) and Uncorrected (U) Correlations for Six Data Sets as a Function of Proportion Selected [P(S)] and k/(k+1)
(See Figures 1 and 2 for Descriptions of the Data Sets)

	k/(k+1)=0.0		k/(k+1)=.25		k/(k+1)=.50		k/(k+1)=.75		k/(k+1)=1.0	
P(S)	C	U	C	U	C	U	C	U	C	U
Data Set 1: r_{yx1}=.39, Linear Regression, Bimodal Conditional Variances										
.20	-33	-49	-21	-43	-26	-53	-17	-59	33	-53
.30	-19	-33	-22	-39	-17	-46	-10	-51	18	-45
.40	-27	-38	-15	-32	-08	-34	-12	-47	11	-43
.50	-22	-34	-10	-27	-10	-34	-09	-41	-07	-47
.60	-24	-33	-22	-35	-19	-37	-09	-36	-08	-42
.70	-24	-33	-22	-33	-17	-33	-19	-39	-02	-30
.80	-18	-24	-19	-28	-13	-26	-11	-27	-09	-28
.90	-09	-13	-10	-16	-08	-16	-05	-16	-09	-21
Data Set 2: r_{yx1}=.42, Flattened Slope, Decreasing Variances										
.20	-42	-52	-49	-60	-48	-61	-33	-58	-61	-81
.30	-40	-50	-44	-55	-34	-51	-66	-78	-79	-89
.40	-34	-43	-36	-48	-49	-61	-64	-75	-49	-70
.50	-34	-43	-44	-55	-57	-67	-37	-55	-54	-70
.60	-30	-39	-32	-44	-37	-50	-38	-53	-46	-63
.70	-21	-30	-25	-36	-28	-41	-37	-51	-40	-55
.80	-15	-23	-15	-25	-21	-33	-20	-34	-24	-39
.90	-10	-15	-08	-14	-07	-16	-12	-22	-11	-22
Data Set 3: r_{yx1}=.36, Increasing Slope, Decreasing Variances										
.20	-20	-53	-11	-49	34	-28	85	-08	167	58
.30	-25	-49	07	-30	34	-20	94	11	137	26
.40	27	00	41	09	52	-02	58	-09	92	00
.50	15	-03	25	-02	26	-13	39	-14	62	-11
.60	13	-03	19	-05	21	-12	46	-01	61	-01
.70	08	-05	15	-05	22	-06	36	-01	47	00
.80	05	-06	10	-06	16	-04	15	-10	23	-08
.90	03	-04	05	-05	08	-04	14	-02	08	-09
Data Set 4: r_{yx1}=.62, Linear Regression, Unimodal Variances										
.20	-44	-69	-28	-63	-06	-57	03	-69	56	-42
.30	-27	-53	-24	-58	06	-44	07	-59	34	-52
.40	-22	-45	-20	-50	-08	-51	07	-50	41	-33
.50	-21	-39	-23	-47	-02	-42	08	-43	19	-47
.60	-14	-30	-16	-39	-03	-37	-01	-44	20	-35
.70	-09	-20	-09	-26	-06	-32	04	-31	07	-37
.80	-10	-19	-08	-20	-06	-24	01	-23	07	-24
.90	-05	-10	-03	-12	-03	-13	01	-12	01	-15
Data Set 5: r_{yx1}=.56, Flattened Slope, Decreasing Variances										
.20	-44	-57	-51	-65	-56	-73	-63	-82	-06	-67
.30	-33	-44	-31	-45	-34	-56	-44	-70	-28	-68
.40	-27	-37	-22	-36	-21	-43	-22	-52	-19	-58
.50	-20	-31	-20	-33	-16	-36	-19	-45	-11	-47
.60	-15	-24	-19	-31	-14	-31	-13	-36	-13	-43
.70	-13	-21	-11	-22	-12	-27	-07	-27	-11	-35
.80	-06	-12	-09	-17	-08	-20	-06	-21	-07	-25
.90	-04	-08	-04	-10	-04	-12	-03	-12	-07	-18
Data Set 6: r_{yx1}=.63, Increasing Slope, Unimodal Variances										
.20	-25	-51	-25	-51	-17	-48	08	-36	36	-29
.30	-11	-33	-11	-33	-08	-35	02	-37	16	-35
.40	-08	-20	-07	-22	-04	-27	04	-29	18	-26
.50	-09	-18	-10	-25	-02	-22	03	-25	18	-19
.60	-08	-16	-07	-19	-02	-19	02	-21	06	-24
.70	-06	-14	-05	-17	-03	-18	02	-17	04	-19
.80	-02	-09	-03	-11	-02	-12	-01	-15	02	-15
.90	-03	-07	-02	-07	-01	-08	00	-08	01	-09

The effect of violating the assumption that selection is based only on x_1 alone can also be seen from the results. An inspection of Table 1 suggests that as the ratio $k/(k + 1)$ decreases, i.e., as selection becomes more a function of x_2 and less of x_1, the values for the corrected correlation tend to decrease. This pattern is exhibited in five of the six data sets. The only major exceptions are found in Data Set 2 where the correction formula is highly negatively biased for $k/(k + 1) = 1$ but somewhat less negatively biased as $k/(k + 1)$ decreases. Thus, in general, if the corrected correlation is negatively biased when selection is based on x_1 alone ($k/(k + 1) = 1$), it will remain negatively biased as selection on x_2 is introduced. In addition, when the corrected correlation is positively biased for selection on x_1 alone, it becomes less positively biased or even negatively biased as x_2 is introduced into the selection model.

It is also of interest to compare the percentage errors for the uncorrected correlation and the corrected correlation for the cases where selection is not based on x_1 alone [$k/(k + 1) \leq .75$]. In all of the data sets except the third, the corrected value is consistently more accurate than the uncorrected correlation. In Data Sets 1, 2, and 5 both the corrected and uncorrected underestimate r_{yx1} for $k/(k + 1) \leq .75$. However, the corrected formula is always more accurate. Further, the superiority of $\hat{r}_{yx1}$ can be considerable, P_C being less than one third the value of P_U in some cases. In Data Sets 4 and 6 the corrected correlation overestimates r_{yx1} to a small degree ($P_C \leq .08$) for $k/(k + 1) = .75$, and underestimates r_{yx1} for $k/(k + 1) \leq .50$. However, in all cases the corrected formula exhibits a considerably smaller percentage error than the uncorrected correlation. In Data Set 3, however, the pattern is reversed, the uncorrected value being more accurate than the corrected value in almost every case. For $k/(k + 1) \leq .75$, the correction formula yields rather large overestimates of r_{yx1}. On the other hand, the uncorrected values underestimate r_{yx1}. Most importantly, the absolute percentage errors are smallest in nearly every case for the uncorrected values.

Why is the corrected correlation less accurate than the uncorrected value for the data set described in Data Set 3? This question can be answered by considering the case where $k/(k + 1) = 1$, i.e., the case where selection is based on x_1 alone. The third data set differs from the other five sets in that the correlation formula produces very substantial overestimates when selection is based on x_1 alone. Although as previously noted, the introduction of x_2 into the selection process decreases this positive bias, $\hat{r}_{yx1}$ still overpredicts even as $k/(k + 1)$ decreases. The large positive bias is due to the particular form of the regression of y on x_1 for the third data set. The regression curve sharply rises as x_1 increases, and the conditional variance of y decreases as x_1 increases. Given this type of conditional distribution and selection based on the highest x_1 scores, it follows that the ratio of residual standard deviations (s_e/s_{e_s}) for the total and selected groups will exceed 1, and the corresponding ratio of the regression slopes (b/b_s) will be less than unity. Thus, the index W given by Equation 2 can greatly exceed 1 and $\hat{r}_{yx1}$ will be highly positively biased.

The results for the third set suggest that the accuracy of the correction formula when selection is not based solely on x_1, i.e., $k/(k + 1) < 1$, can be predicted by considering the special case where selection is completely based on x_1, i.e., $k/(k + 1) = 1$. When the conditional distribution of y given x_1 has a form where $\hat{r}_{yx1}$ is highly positively biased for $k/(k + 1) = 1$, not only is it likely that it will continue to be positively biased for $k/(k + 1) < 1$, but the uncorrected correlations (r_{yx1}) will provide more accurate estimates. In these cases it is disadvantageous to employ the correction formula.

The results can be summarized as follows:

1. The correction formula is not robust to simultaneous violations in the underlying distribution and selection assumptions.
2. Given data where linearity and/or homogeneity of variance are not present, the effect of violating the assumption that selection is based on x_1 alone is in general to decrease the value of the corrected correlation.

3. The corrected correlation can, but need not necessarily, yield more accurate estimates than the uncorrected correlation when both the distribution and selection assumptions are violated. If $\hat{r}_{yx1}$ is not highly positively biased when selection is based on x_1 alone, it will continue to be more accurate than r_{yx1s} when selection is not based on x_1 alone.

Discussion and Conclusion

The analyses for the six data sets and five selection models presented in Table 1 provide useful information concerning the potential accuracy of the correction formula ($\hat{r}_{yx1}$) when neither the underlying distribution nor selection assumptions are satisfied. The generalizability of the results is enhanced by the use of real data sets consisting of actual test score data. The patterns of nonlinearity and/or heteroscedasticity observed in the six data sets are not unique to the present study but may also be encountered by practitioners. For example, the patterns exhibited by Data Sets 2, 3, and 5 were also found in a large scale empirical study by Greener and Osburn (1979). It is clear, however, that the data contained some idiosyncracies. More specifically, some of the trends observed in Table 1 are not monotonic. For example, although it is generally true that the value of the correction formula decreases as selection becomes increasingly a function of x_2, there are few exceptions to this conclusion which can be found in Table 1. However, the data are sufficiently "well-behaved" to support the conclusions drawn.

With respect to the five simulated selection processes, no claim is made that any of these models exactly match real life selection processes. However, they represent useful approximations. The cases in which $k/(k + 1) = .25$, .50, and .75 represent multiple cutoff models where selection is based on some additional variable (x_2) that is correlated with the x_1 variable. The case in which $k/(k + 1) = 0$ represents a situation where selection is based on some additional variable (x_2) that is correlated with the x_1 variable. Finally, for $k/(k + 1) = 1$, selection is solely on the basis of x_1.

The results support two important conclusions. First, it is unreasonable to assume that the correction formula can exactly reproduce or even closely approximate the total group correlation when neither the underlying distribution nor selection assumptions are violated. At best, reasonably small percentage errors in the range of 15% to 20% can be assured only when the degree of selection is quite modest, i.e., only 20% or less of the sample y scores are missing. Further, the data do not support any arguments that violations in the distribution assumption can be "offset" to any practically useful degree by violations in the assumption that selection is solely a function of x_1. Thus, if r_{yx1} is to be estimated with a high degree of accuracy, the correction formula will be inadequate, especially as the proportion of missing y scores increases.

Secondly, in comparing the corrected $\hat{r}_{yx1}$ and uncorrected correlations (r_{yx1s}), situations can occur where $\hat{r}_{yx1}$ is less biased than r_{yx1s}. Thus, in spite of the limitations of the correction formula, it still may be useful to employ $\hat{r}_{yx1}$ for certain types of data. However, it is also possible to find situations where the reverse is true, i.e., r_{yx1s} is less biased than $\hat{r}_{yx1}$. In this case, r_{yx1s} would be the preferred estimator. As examples of the former situation, suppose that $\hat{r}_{yx1}$ were to be negatively biased, i.e., $\hat{r}_{yx1} < r_{yx1}$. When r_{yx1s} is positive, and the variance ratio (s^2_{x1s}/s^2_{x1}) is less than 1, it follows that

$$r_{yx1s} < \hat{r}_{yx1} < r_{yx1} \qquad [8]$$

Thus, even though $\hat{r}_{yx1}$ is negatively biased, it can be a more accurate estimate than r_{yx1s}.

The results for Data Sets 1, 2, and 5 clearly exhibit the inequality pattern described in Equation 8. As a second example of where the correction formula might be preferred, suppose $\hat{r}_{yx1}$ is "slightly" positively biased, whereas r_{yx1s} is "strongly" negatively biased. In such cases it can be argued that r_{yx1} should still be employed. This pattern is evidenced in Data Sets 4 and 6 for the situation where

$k/(k + 1) = .75$. As an example of a case where the uncorrected correlation rather than the corrected correlation should be employed, suppose $\hat{r}_{yx1}$ were to be a highly positively biased estimator of r_{yx1}, i.e., $\hat{r}_{yx1} >> r_{yx1}$. In this case, it is possible for the uncorrected correlation to be negatively biased and to have a smaller degree of absolute bias than the corrected correlation:

$$r_{yx1s} < r_{yx1}$$

$$\hat{r}_{yx1} > r_{yx1}$$

$$|r_{yx1s} - r_{yx1}| < |\hat{r}_{yx1} - r_{yx1}| \qquad [9]$$

When the inequalities described in Equation 9 hold, it is clearly advantageous to estimate the yx_1 correlation using the uncorrected value and to avoid the use of the correction formula. The pattern described by Equation 9 occurs throughout the third data set.

It is of interest to consider the properties of data sets where the inequalities expressed in Equation 9 hold. In other words, what are the characteristics of data sets where the correction formula should not be applied. In general, the form of the conditional distribution of y given x_1 can be characterized in terms of the ratio W described in Equation 2. More specifically, suppose selection were to be based on x_1 alone. The linear regression slopes in the total and selected groups (b, b_s) and the corresponding residual standard deviations (s_e,s_{es}) can be considered. For certain types of distributions, the following will be true when selection is based on the highest scoring x_1 scores:

$$b < b_s \;,\; s_e > s_{es} \;,\; W >> 1 \qquad [10]$$

Thus, the correction formula will substantially overestimate r_{yx1} when selection is based on x_1. Further, the results suggest that the inequalities given by Equation 9 will also hold when selection is not based on x_1 alone. Thus, given a $W >> 1$ distribution form and a selection process not simply based on x_1 alone, the uncorrected correlation can be more accurate than the corrected correlation. An example of this result is clearly seen for the distribution analyzed in Table 3. As previously noted, this distribution has a form where the regression curve is exponential in form, and the variance of y is a decreasing function of x_1.

Given the results of the present study, an answer to the following basic practical question can be attempted: Should the correction formula be applied in real life problems when the underlying distribution and selection assumptions are violated? The only answer that can be given is that "it depends." If prior knowledge is available concerning the form of the yx_1 distribution, the corrected correlation can yield a more accurate estimate than the uncorrected correlation. In general, this prior information may be based upon previous experience with the y and x_1 variables or other variables known to have similar distribution forms. However, when this information is lacking, it may be detrimental to employ $\hat{r}_{yx1}$, since the corrected correlation can be a poorer estimate than the uncorrected correlation. Violations in the underlying assumptions cannot simply be ignored on the assumption that the correction formula will be a robust estimator.

References

CTB/McGraw Hill. *Comprehensive Tests of Basic Skills* (Technical Bulletin No. 2). Monterey CA: Author, 1977.

Greener, J. M., & Osburn, H. G. Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and non-linear distributions. *Edu-*

cational and Psychological Measurement, 1980, *40*, 337–345.

Greener, J. M., & Osburn, H. G. An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurement*, 1979, *3*, 31–41.

Gross, A. L. Relaxing the assumptions underlying corrections for restriction in range. *Educational and Psychological Measurement*, 1982, *42*, 795–801.

Linn, R. L. Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin*, 1968, *69*, 69–73.

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

Novick, M. R., & Thayer, D. T. *An investigation of the accuracy of the Pearson selection formulas* (Research Memorandum RM-69-22). Princeton NJ: Educational Testing Service, 1969.

Sullivan, E. T., Clark, W. W., & Tiegs, E. W. *Test coordinator's handbook and guide to interpretation: Short Form Test of Academic Aptitude*. Monterey CA: CTB/McGraw Hill, 1970.

Author's Address

Send requests for reprints and further information to Alan L. Gross, Graduate Center, The City University of New York, 33 West 42 Street, New York NY 10036, U.S.A.