DOCUMENT RESUME

ED 387 523 TM 023 817

AUTHOR Lewis, Charles; Willingham, Warren W.
TITLE The Effects of Sample Restriction on Gender Differences.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-13
PUB DATE Mar 95
NOTE 68p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Age Differences; *Educational Assessment; Models; *Sampling; *Scores; *Selection; *Sex Differences; Simulation; Testing
IDENTIFIERS *Restriction of Range; *Restrictive Procedures

ABSTRACT

As strongly suggested by recent work, patterns of gender difference can change because of changes in the selectivity of the sample itself. This is a statistical influence connected with the distributions of female and male scores, rather than a substantive influence related to demographic characteristics of the sample such as age or ethnicity. It is, nonetheless, an important influence because gender differences that are partly statistical in origin can easily confuse possible implications regarding education and assessment. This report proposed a general model to account for the effects of sample restriction on gender differences. Simulations showed the model to be quite accurate in reproducing standard mean differences and other statistics in a restricted sample. Three primary contributing factors were identified: the range-restricting effects of sample selection, differential variability of female and male scores in the original sample, and the representation of females and males in the restricted sample. A test with actual data showed reasonably good consistency between trends predicted by the model and trends in gender differences that have been widely observed in advanced tests administered to select samples. (Contains 29 references, 9 tables, and 10 figures.) (Author)

RESEARCH REPORT

THE EFFECTS OF SAMPLE RESTRICTION ON GENDER DIFFERENCES

Charles Lewis
Warren W. Willingham

Educational Testing Service
Princeton, New Jersey
March 1995

The Effects of Sample Restriction on Gender Differences

Charles Lewis and Warren W. Willingham*
Educational Testing Service

To appear in Willingham, W. W. and Cole, N. S., Gender and Fair Assessment.
Princeton, NJ: Educational Testing Service.

*Authors are listed alphabetically.

Abstract

As strongly suggested by recent work, patterns of gender difference can change because of changes in the selectivity of the sample itself. This is a statistical influence connected with the distributions of female and male scores, rather than a substantive influence related to demographic characteristics of the sample such as age or ethnicity. It is, nonetheless, an important influence because gender differences that are partly statistical in origin can easily confuse possible implications regarding education and assessment.

This report proposes a general model to account for the effects of sample restriction on gender differences. Simulations showed the model to be quite accurate in reproducing standard mean differences and other statistics in a restricted sample. Three primary contributing factors were identified: the range-restricting effects of sample selection, differential variability of female and male scores in the original sample, and the representation of females and males in the restricted sample. A test with actual data showed reasonably good consistency between trends predicted by the model and trends in gender differences that have been widely observed in advanced tests administered to select samples.

Acknowledgements

This report is part of a larger project directed by Warren Willingham and Nancy Cole on gender differences and similarities in achievement and implications for fair assessment. We are grateful to Nancy Cole for useful conversations about the complex topic in which the particular problem here addressed is imbedded. We wish to thank a number of reviewers: the project advisory committee (Pat Campbell, Richard Duran, Robert Linn, Lorrie Shepard, Richard Snow, Floraline Stevens, Carol Tittle, and Janice Weinman); ETS colleagues: Sam Messick, Nancy Petersen, Larry Stricker, and Howard Wainer; and also, Larry Hedges for his careful review. We are especially grateful to several colleagues for their assistance: to Judy Pollack for help with the analyses, to Linda Johnson for several contributions to the project and to Joan Stoeckel and Carol Crowley for help with the manuscript.

The Effects of Sample Restriction on Gender Differences

This report addresses the following question: What is the effect of sample restriction on observed gender differences? It may be useful to draw some distinctions as to what we mean by sample and restriction. We use the term population mainly to refer to age or grade cohorts which, for our purposes, are unrestricted. We refer to samples intended to represent such populations as national or representative samples. We are here concerned with restricted samples that clearly do not represent national cohorts. Sample restriction can occur in a variety of ways, but we refer primarily to the fact that tests are often administered to a selected (often self-selected) group of individuals who, on average, score higher than would a representative sample at the same age level.

One view of the problem is empirical. Why is the standard mean difference, *D*, in selected samples more likely to favor males, as compared with *D*s for similar tests in unselected, representative populations? (In computing *D*, we subtract the male mean from the female mean.) Elsewhere we have compared *D*s in large selected and unselected populations of students for a number of tests administered at the same grade level (Willingham & Cole, in preparation). Results for individual tests varied, but in each of nine categories of tests where such a comparison was available, the average *D* was more negative in the restricted, more able group. This drift in *D*, favoring males, averaged -.18 for similar types of tests at the same grade level. The character of the sample would seem to be a natural suspect in trying to account for this systematic change when construct and cohort appear to be generally comparable.

Another view of the problem is analytical. It is clear that selection of students from one tail of a distribution can change the balance of females and males--certainly one important indicator of gender difference. With normal distributions and equal standard deviations, explicit selection must yield for the group with the higher mean a more favorable *F/M* (female/male) ratio in the upper tail than exists in the original group. The higher the cut, the greater that effect. If the standard deviations differ, the *F/M* ratio will change with selection even when the mean difference is zero. Figure A shows that such an effect will be heightened if the group with the higher mean is also more variable. It seems safe to assume that *D* will be affected as well (both within the shaded tail area and at any particular percentile score for females and males), but it is less obvious how.

Insert Figure A

Our objective, then, is to improve understanding of the purely statistical effects of sample restriction on indicators of gender difference. That is, how

do various conditions of sample restriction likely influence gender differences, independent of test characteristics or changes in the cohort due to learning or other intervening experience? Since these factors are not normally independent, a secondary interest is the extent to which such effects appear to account for changes in the actual pattern of *D*s commonly observed in tests administered to selected samples of students.

Background of the Problem

Critical to any detailed understanding of subgroup differences is the fact that individuals vary widely within each subgroup and that *D*, which only characterizes the mean difference, is a limited and sometimes misleading indicator. It is reasonable to argue, as have Snow and Ennis (1992), that one's concern should be to understand group differences at all relevant points on the score scale--partly because the differences may vary substantially at different points, and partly because the educational meaning and implication of those differences may also be substantially different. It is one thing to say that, in a particular grade, girls are somewhat better readers on average. It is quite another thing to say that, in that same grade, twice as many boys as girls are essentially non-readers.

Over the years, practical interests and educational implications have been reflected in much of the literature on specific types of gender difference--rare talents, risk factors, affective characteristics, physical abnormalities, and so on (Obler & Fein, 1988; Singer, Westphal, & Niswander, 1968). Historically, evidence regarding such variations has been used to speculate about human and subgroup variation generally, though the connection is often tenuous at best. Such unusual variations imply the possibility--even the likelihood--of distributional differences, though most systematic statistical studies have focused on the mean. The most obvious and useful complement to the mean in understanding the range and character of subgroup differences on a particular characteristic is, of course, the standard deviation. Like the mean, the standard deviation is a somewhat oversimplifying descriptive statistic. It is, nonetheless, a significant addition in trying to unravel a complex topic.

There is a long history to the interest in gender differences in variability. Key aspects of that literature were cited earlier (see especially Feingold, 1993; Maccoby & Jacklin, 1974; McNemar & Terman, 1936; Noddings, 1992). Nonetheless, distributional aspects of subgroup differences have received limited systematic attention. In fact, the most significant advances in understanding gender differences in variability and the important role variability plays in sample restriction are quite recent.

Two Important Lines of Research

In examining performance of students in a classroom or school grade, it is often natural to focus on comparisons at a particular level of proficiency; for example, whether pupils in the bottom decile are showing improved skills, year to year. In that spirit, several researchers at the University of Iowa have used a statistic which we call here D_P , or "standard percentile difference" (Becker & Forsyth, 1990; Cleary, 1992; Han, Cleary, & Rakaskietisak, 1992; Martin & Hoover, 1987). This statistic is a derivative of the familiar *D*. It simply standardizes the score difference at a particular percentile level for two subgroup distributions (e.g., female 90th percentile minus male 90th percentile) by dividing the difference by the usual average or pooled within-group standard deviation for the total group.

Cleary (1992) has made the most extensive use of this statistic, and in so doing, usefully informed the topic at hand. Noting in her large study of gender differences on aptitude and achievement tests that males were more variable on "almost all tests," Cleary used D_P to examine gender differences at three points on the female and male distributions: the 10th, 50th, and 90th percentile points. Overall, she reported small between-group differences, generally favoring males, but with the greatest difference favoring males at the upper percentile levels. Given the tendency to greater male variability, that outcome is illustrated in Figure A. With normal distributions, it is true but not obvious that a greater male *SD* results in a greater female-to-male difference at the 90th than at the 50th percentile.

Cleary's results also showed consistently larger gender differences favoring males on tests administered to select as opposed to representative groups. Taken together, these results suggest that in restricted samples of higher scoring students, gender differences may be naturally inclined to favor males because of differential variability, irrespective of other factors such as test content. Cleary did not examine that possibility, though she signaled the pervasiveness of distributional differences and illustrated why they are important.

In another significant line of work, other writers have addressed more directly the joint role of means and standard deviations in describing gender differences (Feingold, 1992a, 1992b, 1993, 1995; Hedges & Friedman, 1993a, 1993b). Noting the almost exclusive use of *D* in meta analyses of gender differences, Feingold (1992a) questioned whether evidence concerning the relative size of *SD*s for females and males would support the reasonableness of that practice. He assembled extensive normative data on variability from several test batteries in order to test this assumption. His first conclusion, like that of Maccoby and Jacklin (1974), was that males are more variable in mathematical

and spatial abilities but not in verbal ability. Second, Feingold provided what he called subjective estimates of the effect of differences in both mean *and* standard deviation. This was accomplished by determining what *D* would be required at the mean in order to produce the observed female-to-male ratio in the tails if there were no difference in standard deviations. The same principles apply, of course, in the upper and lower tail, but our focus is on the upper tail. He found that the gender difference, so characterized, often gave a different impression of magnitude, compared to that based upon *D* alone. On this basis he urged the consideration of both mean and *SD* in order to understand gender differences at different performance levels. See Feingold (1995) for further discussion of this approach.

Hedges and Friedman (1993a) disputed Feingold's estimates of effects in the tails and demonstrated that, with some assumptions, it was possible to estimate a precise value for *D* within the tail. Separately, they also derived estimates of the ratio of males to females in the tail. There ensued a scholarly exchange in which the parties maintained some disagreements, but it became evident that their two estimates of *D* represent two views of the same picture.

This line of work is important because it represents the first systematic effort to integrate the effects of mean and standard deviation conceptually and analytically. It resulted in Hedges and Friedman's (1993b) useful derivations upon which the current work builds. All of these authors clearly agreed on the importance of examining female-to-male representation in the tail. Feingold (1992) was the first to urge formal integration of that metric into the analysis of gender differences.

Extensions of Recent Work

Each of these two lines of work concerns the interpretation of gender differences in a segment (i.e., tail) of a distribution. One refers to the standard gender difference between means *within* the shaded tail in Figure A. The other refers to gender difference between scores *at* a particular percentile. We focus on the upper tail, though the principles are the same at both ends of the distribution. Both approaches carry some suggestion as to what gender difference one might expect to observe in a different sample, somewhat similarly restricted, though neither actually addresses that question nor intended to. Both approaches also suggest what role differences in variability can play, although some extensions are useful to make the relationships clearer and to show how the two lines of work are related.

First, a word on notation. We use the subscript *o* to designate the *original* group, *r* to designate a *restricted* sample. These may refer either to different samples altogether or to a total sample versus a restricted portion of that sample. Thus,

D_o is the standard mean difference in an original unrestricted group;

D_r is the standard mean difference in a restricted sample.

D_p refers in general to a standard percentile difference which is normally designated with the pertinent percentile subscript, for example:

D_{90} is the standard score difference between female and male scores at their respective 90th percentiles.

Hedges and Friedman (1993a) present formulas that provide estimates of *N*, mean, and *SD* for females and males in the restricted group. From these analytic expressions, explicit and useful as they are, it is difficult to see the relationships of particular interest; for example, how D_r in the restricted group can be expected to vary with the *SD* ratio (*SDR*) and D_o in the original group and with the proportion selected, P_r . Hedges and Friedman provide estimates of D_r and F/M for Feingold's data--which was their main purpose--but relationships among these measures are not easily recognizable.

In the interest of further clarifying, we have extended the analytic work of Hedges and Friedman in order to develop expressions for D_r and F/M as a function of *SDR* and D_o in the total group. Selected values are shown in Table A, and the relationships are plotted in Figures B and C. These curves may be closely approximated as follows. (Note that the curves in the figures are actual values, not the approximations.) When $P_r = .10$,

$$D_r \doteq .41D_o + 1.67\ln(SDR) \quad (1)$$

$$\ln(F/M) \doteq 1.78D_o + 2.29\ln(SDR) \quad (2)$$

Insert Table A and Figures B, C

The standard percentile difference (D_P) used by Cleary (1992) and others poses similar interpretive limitations. It is apparent in Figure A that *D* at the 90th percentile varies with the original standard mean difference, D_o , but will

also be affected by unequal female and male variance in the original group. The main idea of the statistic is to index such an effect, though the exact nature of the relationship has not been clarified. It can be shown that D_{90} is approximated by

$$D_{90} \doteq D_o + (1.28)\ln(SDR)$$

In general,

$$D_P \doteq D_o + Z_P\ln(SDR) \qquad (3)$$

for *SDR*s not too different from 1.0. Illustrative values of D_{90} are shown in Table B. The preceding equations and illustrative data indicate several useful relationships, some of which seem obvious, others less so. We examine two aspects of gender difference and similarity in turn: the standard difference and the female-to-male ratio.

Insert Table B

Standard Difference. The first line of Table B confirms that, if females and males do not differ as to variability (*SDR* = 1.0), the Standard Percentile Difference, D_P, is always the same as D_o at the mean. On the other hand, Figure B shows that with no gender difference in variability, the Standard Mean Difference within the tail (D_r) is always smaller in absolute value than D_o at the mean. Clearly, the variance of scores is reduced in the tail, but any overall difference between female and male means is even more reduced.

With both coefficients, D_P and D_r, there can be a consequential effect of a gender difference in variability in the original group. Differential variability has a generally similar effect on D_P and D_r at the 90th percentile. With normal distributions, a drop in *SDR* from 1.00 to .85 will tend to lower these two versions of *D* in this upper tail by some .21 or .27 points, respectively, on a standard scale; that is, move *D* in a negative direction. As Tables A and B suggest, and Equations (1) and (3) confirm, the effect of unequal variability is more or less constant at all levels of original *D* within this range.

On the other hand, D_P and D_r are not similarly affected by unequal variability at different levels of sample restriction. It is obvious from the nature of the statistic that, with symmetric distributions, D_{50} will not differ from D_o because of unequal female-to-male variability. Equation 3 confirms that. On the other hand, Table A shows that unequal standard deviations affect D_r almost as much in the mid-range as in the tail.

Female-to-male ratio. The ratio of females to males (F/M)in a restricted sample is a second important index of gender difference or similarity. Comparing D_P and F/M is a moot issue since, by definition, D_P refers to female and male groups of equal size. But D_r assumes, in this case, that females and males are selected on the basis of a single cut score, which means that F/M in the restricted group will vary and will surely be affected by *SDR*.

That relationship is illustrated in Figure C. First, assuming females and males are equally variable, a restricted sample will contain more females if the original D_o was positive, more males if D_o was negative. As we have noted in previous discussion, and as is illustrated in Equation (2) and Table A, the farther out into the tail one looks, the greater this differential representation. Figure C looks very much like Figure B. *SDR* and D_o have similar, independent effects on the two indices of gender difference represented in these two figures. In both figures differential variability has an essentially constant effect across the observed range of D_o. Also, the effects can be consequential. When there is no gender difference at the mean, a *SDR* of .85 versus 1.00 results in a F/M ratio of about .70 rather than 1.00. Finally, the influence of *SDR* is magnified in the tail. As Table A shows, when a sample is restricted in mid-range, *SDR* has essentially no effect on F/M in the resulting sample.

What can we conclude from these analyses concerning the likely effects of differential variability on observed gender differences in different parts of a distribution? Three things, in the main. First, the effects are quite orderly, approximately proportional to the size of ln (*SDR*) within the range studied, and essentially constant for different values of D_o . Second, as previous writers have argued, it is clear that the effects of differential variability can be consequential. This is particularly true in the tail where these effects become progressively larger. Third, variation in *SDR* has remarkably similar effects on both D and F/M-- similar in character as well as direction. The effect of *SDR* is thus compounded to the extent that D and F/M are viewed as two distinct and important aspects of gender similarity and difference.

The foregoing analyses concern the description of gender differences *within* a sample. They pertain to a subgroup of that sample restricted on the basis of explicit selection on the measure of interest. Some of the most interesting and important effects of selection on group differences occur, however, in other situations that engage additional complications. We move now to the more general problem of sample restriction.

A Sample Restriction Model

How does the more general problem differ from the situation we have been discussing? First, there are a wide variety of naturally occurring restricted samples of interest, and for the most part, they do not come about through explicit selection as implied in Figure A and the types of analyses just described. They tend to result from far more complex implicit selection processes in which several factors may come into play, and often there are no directly comparable data available on the original unrestricted sample.

Suppose that we have test results for a selected group of students, and we would like to know how sample restriction may have affected the observed pattern of gender difference. Scores are available only for the selected group. For example, students do not decide to take the most advanced mathematics achievement test, (Math II) offered by the College Board on the basis of their score on that test. They decide rather on the basis of their interest and preparation and whether their choice and performance will likely advance their admissions cause. Such factors have an indirect implicit relationship with score level and resulting gender difference in the restricted sample of Math II examinees.

Also, in the explicit selection procedures previously discussed, the number of females and males in a restricted group comes about inevitably as a necessary result of a particular selection score and the characteristics of the original group. In many restricted groups, however, the balance of females and males is an active and often critical aspect of the sample restriction process itself. This is because many factors are involved in most selection situations--factors that are specifically gender related, such as differential interest patterns. Finally, the most appropriate standard for evaluating the effects of sample restriction on gender differences in performance is not altogether obvious. As we shall see, there may be two standards, equally useful. The foregoing considerations suggest three key issues that need to be taken into account in studying the effects of sample restriction on gender similarity and difference. These are: implicit selection, gender-related sample restriction, and what standard to use in evaluating differences. The following discussion of these issues lays the groundwork for a description of the analytic model we propose.

Three Key Issues

Implicit selection. One may be interested in gender difference and similarity in a wide variety of restricted samples. For example, there are students who comprise the college-going pool, adults who follow a particular profession, pupils

who require special instruction, applicants to a graduate program, members of a college freshman class, examinees who take a subject test for advanced placement, and so on. These groups can be viewed as selected samples from a general population such as an age or grade cohort. The groups will vary in many ways, as will the females and males within each group--collectively and individually.

Such restricted samples are seldom formed simply through explicit selection on a particular measure. In fact, a measure of interest--say, an admissions test taken by the restricted group of students who plan to go to college--is normally not even administered to an unrestricted representative national sample of 12th graders. "Restricted" here means the net effect of self-initiated, institutional, and circumstantial factors that result in the particular selected sample of individuals who take the test. In order to create an analytic model of the problem, it is useful to think of a restricted sample as having come about through explicit selection on some hypothetical composite, X, based on all variables that determine who ends up in the sample. Any given measure of interest, Y, will have some correlation, r_{XY}, with that composite. There are implicit selection effects on variable Y because Y is related to the variables that make up the composite; that is, the factors that actually determine who takes the test. Such a hypothetical relationship is illustrated in Figure D.

Insert Figure D

Explicit selection on such a composite results in implicit selection on variable Y as indicated by the dashed ogive in the upper panel. The cross-hatched areas in the lower panel illustrate that due to implicit selection, most but not all individuals with high Y scores self-select into the restricted group, as do even a few with low Y scores. The variable of interest, Y, may have a characteristic gender difference, as suggested by the two distributions in the upper panel. Also, there is presumably a characteristic gender difference on composite X, which reflects the fact that all things taken into account, either females or males "score" higher on the hypothetical composite because they are more likely to fall into the restricted sample. In view of that likely difference, one could show two ogives, one for females and one for males. Such ogives represent the differential likelihood that females and males at various score levels on Y will be in the restricted sample. Thus, gender difference in a restricted sample results partly from D_o in the original sample and partly from the measure's weight (correlation) as an implicit variable in a selection process that, for many reasons, results in a given number of females and males in the restricted sample. This brings us to the second key issue.

Gender-related sample restriction. Observed gender differences in a restricted sample can be influenced markedly by gender-related aspects of the selection process. Pursuing the previous example, the number of women and men who elect to take a Math II College Board Achievement Test will be influenced by various factors: attitudes about the test, who schools and teachers encourage to take the test, and so on. As we shall see, such factors can have large effects on observed gender differences in such restricted samples of test takers. The possible contrasting outcomes are more easily illustrated with explicit selection, as shown in Figure E. The principle is the same with implicit selection.

Insert Figure E

The female and male distributions in Figure E represent two groups of equal size that are separated at the mean by a standardized value of -.20, a "small" difference. The arrows illustrate reciprocal outcomes of two opposing ways that a sample might be restricted. The solid vertical line represents "gender-blind" selection resulting from a single cut at the 90th percentile of the total distribution. There is very little difference in the mean performance of females and males in the resulting restricted samples (D_r = -.08), but the F/M ratio has dropped to .70 as a result of unequal numbers of females and males in the tail (see Table A). The dashed lines represent a selection process based on separate cuts for females and males--at the 90th percentile of each distribution. These cuts result in equal numbers of females and males (in the two shaded tails), but that balance comes at the expense of a substantially larger D_r of -.48. The latter estimate comes from Table G and analyses to follow.

As we shall see, the effects under implicit selection are similar, though moderated if the measure in question has a low correlation with the selection process. If the variable is heavily weighted in selection, the trade-off stakes will be high as suggested by the contrasting results in Figure E. How does gender get involved in sample restriction so as to influence such reciprocal outcomes? Differential attitudes and interests will frequently play a role in individual decisions. One obvious example is the well-documented difference between young women and men in the value they perceive in advanced mathematics (Chipman & Wilson, 1985). Females and males may differ--probably in different ways--on important characteristics, preparation, or accomplishments that bear on the restriction process. Membership in the restricted group may also be influenced by unequal opportunity, discrimination, or some deliberate action with benign intentions, such as affirmative recruiting or a college's effort to achieve some desired balance of females and males in its freshman class.

As a result of all such factors operating on an unrestricted group, some proportion of the females (P_F) and some proportion of the males (P_M) will make it to the restricted group. Thus, depending upon how it comes about, gender-related sample restriction (that is, P_F and P_M) can be expected to have reciprocal effects on two key indicators of gender difference: direct determination of F/M, and indirect influence on D_r through the rippling effects of implicit selection.

It is important to appreciate that this reciprocal trade-off between D and F/M can be expected to hold and is strictly predictable only for a given set of scores, within a given sample, where females and males are similarly selected. F/M and D would not be limited by their statistical relationship if any of these three assumptions change.

What standard? In examining gender or any other subgroup difference in a restricted sample, what standard of comparison is most appropriate? The standard conventionally used is some variation on the average standard deviation for the groups being compared. But there is an alternative that can often be informative, and may actually be more appropriate. Again, Figure E can serve to illustrate.

Assume we have a restricted sample consisting of the two shaded areas. As indicated, the standard difference between females and males is equal to -.48. But if one wishes to evaluate differential mean performance of females and males on the original proficiency scale, then it would be more appropriate to use a modified coefficient D', wherein the denominator would be based upon the within-gender *SD* for the total *unselected* group. Similar logic might apply if one wished to compare gender difference in a restricted group with, say, differences in a general population of 12th graders. The reasoning stems from this natural question: To what extent might an observed D_r be larger than the original D_o, simply because there is less variation among scores in the restricted group? It can be shown that, if we select those females and males above the P^{th} percentiles of their respective Normal distributions, so that $F/M = 1.0$ (as in Figure E), D' may be approximated by

$$D' \doteq D_o + \bar{Z}_P \ln(SDR) \tag{4}$$

for *SDR*s not too different from 1.0. (Here $\bar{Z}_P$ denotes the mean of the tail of a standard Normal distribution above the P^{th} percentile.) As illustrated in the example in Figure E, with Normal distributions and equal female and male standard deviations, the actual mean difference between the shaded female and male upper tails will be the same as the mean difference for the total groups.

Thus if the original standard for comparison is employed, then D' is equal to D_o whenever female and male score variability are the same. In this case, all of the apparent increase in gender difference, from -.20 to -.48, is due to range restriction. Notice that the relationship of D' to D_o in Equation 4 is parallel to the relationship of D_P to D_o in Equation 3.

Since variability in a restricted group will be smaller than variability in the original unrestricted group, D' will be smaller than D_r in absolute size. A comparison of the two provides a convenient index of the extent to which an apparent change in gender difference is due to range restriction alone. We provide for that comparison in the analytic relationships (Table C) and empirical illustrations (Tables F through H) to follow.

In using the original scale of proficiency, one might reasonably question whether a difference of given magnitude means the same thing or has the same consequences at different points on the scale. Those are reasonable questions that deserve attention when considering such differences at different scale points. It is not clear, however, that such considerations argue for using only a conventional standard that exaggerates the difference and varies with the character of the sample restriction. An argument for the conventional standard might be the likelihood that, in time if not initially, judgments about differences and consequences of differences play out in relation to the amount of variation among the individuals in the restricted sample. It is a debatable point worth consideration.

Analytic Characteristics

In the model depicted in Figure D, we start with distributions of scores on Y for females and males which describe a pattern of gender difference in the original group. Selection on the hypothetical composite X will produce restricted female and male samples of varying size and shape, depending upon the proportion selected and the extent to which X is correlated with Y. The problem is to estimate N_r , mean, and SD_r for each of the selected samples of females and males in order to ascertain the pattern of gender difference in the restricted group. The statistics of interest are summarized in Table C.

Insert Table C

Notation in the first column of Table C refers to characteristics of scores in the original unrestricted group. These statistics, used separately for females and

males, provide the input information that is needed in order to estimate N, mean and SD in the restricted group. The second column contains derivative statistics defined earlier that are useful in characterizing gender differences in the original group. The last column refers to the restricted group. It shows those same statistics plus one other, D', which is D_r corrected for range restriction.

As indicated by the note in Table C, we have special interest in some of these statistics. In order to estimate the effects of sample restriction on gender differences, we need each of the five paired items of information in column one. If starting with an original unrestricted group and predicting forward, the first three are available. The fourth statistic, P, is also available, either because we know how many females and males there are in the restricted group of interest, or we wish to examine how D_r varies with different assumptions about the proportion of females and males selected. The proportion selected is one of two critical aspects of the sample restriction process and, for that reason, P is a variable of special interest.

The fifth input statistic, r (overall, as well as, for females and males separately), is also a key piece of information, partly because it is always unknown. Since r represents the relationship of Y to the outcome of the selection process, it can be represented by r_{XY} or r_{bis}, the biserial correlation of Y with membership in the restricted group. In theory these two are identical. In real situations one is more likely to use r_{bis} since X is usually only a hypothetical variable. All empirical estimates of r_{XY} are here based on r_{bis}. Estimates of r can be based on information available on similar measures, and as we shall see, r can be estimated from simulations based on plausible estimates. It is also a key statistic because it represents the second critical characteristic of the restriction process, namely, how much weight Y has in selection. That weight is reflected directly in r_{bis}.

In the second column of Table C, both D_o and SDR carry special interest. D_o is the primary baseline; for example, with a given D_o in the unrestricted group, what D_r would be expected if such and such are assumed. The special interest in SDR is because it represents a characteristic of the original distributions of scores that can have a large impact on restriction effects. The starred statistics in the first two columns are the ones most useful in conceptualizing and illustrating graphically the effects of sample restriction; that is, D_o, SDR, r_{XY}, P, and the gender difference in P.

The last column of Table C lists the descriptive outcome statistics of interest. The conventional measure of standard difference, D_r, is paired with the comparable measure, D', which is corrected for range restriction. The second important aspect of gender similarity or difference is F/M_r. It is complementary

to D_r because it denotes relative female-to-male representation rather than relative level of performance. F/M is a curious statistic because, from different legitimate perspectives, it can be viewed as both an input and an output measure--as input in examining the possible consequences of sample restriction, as output in examining the possible costs and benefits of reciprocal patterns of representation versus performance level for females and males in a restricted group. We take up now the equations necessary to express the relationships among these statistics.

Definitions and Derivations

First, we need to review notation. There are three group membership variables: F for female students, M for male students and r for those students in the restricted group. For simplicity, let us suppose that students in the restricted group can be defined as those for whom $X>C$ for some cut score C on the X-scale. If we want to consider the possibility that different cut scores are appropriate for female students and for male students, we indicate that with subscripts: C_F and C_M.

As previously noted, we use P to specify the proportion of students belonging to some group. Thus P_r denotes the proportion of students in the restricted group. An important additional consideration is our use of conditional notation. For instance, $P_{r|F}$ refers to the proportion of female students who are in the restricted group.

We use the shorthand notation $\bar{Y}$ for the mean of Y, or $E(Y)$. For conditional means, such as the mean of Y for female students in the restricted group, or $E(Y|r,F)$, we will use $\bar{Y}_{r,F}$. We denote the standard deviation of a variable by SD, with, for instance, $SD_{X|M}$ referring to the standard deviation of X for male students. Of special interest will be $r_{XY|F}$ and $r_{XY|M}$, the correlations between X and Y for female and male students, respectively.

Putting our notation to use, we define two standardizations of the cut score(s):

$$Z_{C|F}=\frac{C_F-\bar{X}_F}{SD_{X|F}}$$

and

$$Z_{C|M}=\frac{C_M-\bar{X}_M}{SD_{X|M}}$$

for female and male students, respectively.

In order to derive expressions for the proportions, as well as means and standard deviations for Y, of female and male students in the restricted group, we introduce two pairs of assumptions. The first is that the scores for female students and for male students in the original group on the hypothetical composite X each have Normal distributions. The second is that the regressions of Y on X for female students and for male students in the original group are each linear with homogeneous residual standard deviations. Note that we do not assume X to have the same distribution for female and male students, nor do we assume that the regressions of Y on X (or the residual standard deviations) are the same for the two groups. Also note that the framework adopted here for discussing the effects of restriction corresponds to that used for deriving the Pearson-Lawley equations in order to estimate correlations corrected for restriction of range (see, for instance, Gulliksen, 1950, p. 128-143).

Combining the first pair of assumptions with the earlier assumption that the students in the restricted group are those whose scores on X exceed the cut score(s) C, we may immediately write expressions for the proportions of female and male students in the restricted group:

$$P_{r|F}=1-\Phi(Z_{C|F})$$

and

$$P_{r|M}=1-\Phi(Z_{C|M})$$

where $\Phi(.)$ denotes the standard Normal cumulative distribution function. Of course, in practice, since X is hypothetical, it is unlikely one would know the cut scores on X. Instead, it is more likely that one would know the proportions of female and male students in the restricted group. In such a case, we would solve these expressions for $Z_{C|F}$ and $Z_{C|M}$, identifying the standard normal deviates giving Normal tail areas equal to the known proportions.

To obtain the means and standard deviations for Y in the restricted group, we first need expressions for the mean and standard deviation of a standard Normal random variable restricted to be greater than Z_C. (The resulting distribution is sometimes referred to as a truncated Normal.) These are given by (Johnson & Kotz, 1970, p. 81-83) as:

$$\bar{Z}_r=\frac{\phi(Z_C)}{1-\Phi(Z_C)}$$

and

$$SD_{Z|r}=\sqrt{1+Z_C\bar{Z}_r-\bar{Z}_r^2}$$

respectively, where $\phi(.)$ denotes the standard Normal density function. The mean of Y for female students with a standardized value of X equal to $Z_{X|F}$ is given by the regression equation

$$\hat{Y}_F=\bar{Y}_F+(r_{XY|F}SD_{Y|F})Z_{X|F}$$

We obtain the mean of Y for female students in the restricted group by taking the mean of this expression over all X above the cut score, which is equivalent to substituting $\bar{Z}_{r,F}$ for $Z_{X|F}$ in the regression equation:

$$\bar{Y}_{r,F}=\bar{Y}_F+(r_{XY|F}SD_{Y|F})\bar{Z}_{r,F}$$

The corresponding expression for the male students in the restricted group is

$$\bar{Y}_{r,M}=\bar{Y}_M+(r_{XY|M}SD_{Y|M})\bar{Z}_{r,M}$$

Turning now to the standard deviations of Y for female and male students in the restricted group, there are two sets of components to consider. The first of these are the residuals around the regression lines. The residual standard deviations of Y given X have the form

$$SD_{Y|X,F}=SD_{Y|F}\sqrt{1-r^2_{XY|F}}$$

for the female students and

$$SD_{Y|X,M}=SD_{Y|M}\sqrt{1-r^2_{XY|M}}$$

for the male students. The second set of the components are the mean values of Y given X. The standard deviations of these for X above the cut score(s) may be obtained by multiplying the standard deviations of $Z_{X|F}$ and $Z_{X|M}$ in the restricted group by their coefficients (assuming these are positive) in the regression equations for the female and male students:

$$SD_{\hat{Y}|r,F}=(r_{XY|F}SD_{Y|F})SD_{Z|r,F}$$

and

$$SD_{\hat{Y}|r,M}=(r_{XY|M}SD_{Y|M})SD_{Z|r,M}$$

The variances of Y for female and male students in the restricted group are each given as the sum of the variances of these two components, so the standard deviations may be written for the female students as

$$SD_{Y|r,F}=\sqrt{SD^2_{Y|X,F}+SD^2_{\hat{Y}|r,F}}$$
$$=SD_{Y|F}\sqrt{1-r^2_{XY|F}(1-SD^2_{Z|r,F})}$$

and for the male students as

$$SD_{Y|r,M}=\sqrt{SD^2_{Y|X,M}+SD^2_{\hat{Y}|r,M}}$$
$$=SD_{Y|M}\sqrt{1-r^2_{XY|M}(1-SD^2_{Z|r,M})}$$

With the expressions we have provided for proportions, means and standard deviations of Y in the restricted group, we can compute F/M ratios, standard mean differences and standard deviation ratios once we know means and standard deviations of Y in the original group, proportions of female and male students selected and the correlations with the selection variable for female and male students.

Operating Characteristics

How does the proposed model work with real data? There are two main questions. One concerns the extent to which distributional assumptions of the model are likely to be met. Another concerns the extent to which the effects of sample restriction that are predicted by the model are reproduced empirically. To examine these questions, we have utilized the 1992 followup sample from the National Education Longitudinal Study (NELS). The NELS data base is useful because it includes four test scores (reading, mathematics, science, and history) and transcript-based high school grade information. Also, this large representative sample of 12th grade students can be restricted to simulate realistically the types of changes in gender patterns that may occur in more selected samples of test-takers.

Distributional assumptions. The initial basic assumption of the model is that the distribution of scores in a restricted sample is generated by a Normal ogive function such as that represented in Figure D. That is, the probability of an individual female or male student falling into the restricted group increases

regularly with increasing scores on the variable of interest. This relationship is illustrated in Figure F, which shows what proportion of 12th grade seniors at different achievement levels fall into the restricted sample of students taking a college admissions test. The data are based upon responses of students in the NELS sample to questions about admissions tests. Figure F shows what proportion of students at each high school grade level or NELS composite score level (an equally weighted average of the four tests) said that they had taken either the ACT or SAT.

Insert Figure F

The empirical plots in Figure F show a function that is similar in slope and shape for females and males on the NELS test composite and the high school average. One difference readily apparent in the figure is that, through most of the range of test scores, the proportion of females taking an admissions test was some .10 to .20 higher than that of males--a difference not observed with respect to high school average. The most parsimonious accounting would suggest that this difference is simply the net result of women earning somewhat higher school grades and slightly lower test scores overall in this sample (D = .31 and -.09, respectively) but being somewhat more likely to take college admissions tests than are men (F/M = 1.18).

Regardless of the underlying reasons why individual students take tests and go to college, the curves indicate that probability of membership in the restricted group of test-takers does increase with achievement level as assumed by the model. While this would not appear to be an unreasonable assumption, these data do not speak to its generalizability.

A second basic assumption of the model--an assumption shared by the derivations of Hedges and Friedman (1993b)--is that the score distributions of interest are Normal. Whereas the previous assumption seems reasonably safe, this assumption of normality may be more problematic because the shapes of distributions can easily vary. We undertook a fairly stringent test of the model's robustness to violations of this assumption as follows. The NELS composite, based on all four tests, was used as the selection variable. A restricted sample was selected as the top 10% of scores on this composite. Using r_{bis} in the equations previously specified, mean and SD were predicted for females and males, as well as the value of D_r, for each individual test score within the restricted sample.

Details are shown in the top section of Table D, where the results are unambiguous. All means in the restricted group were predicted with considerable accuracy. All *SDs* were overpredicted, typically by a factor of two or more. All values of D_r were correspondingly underpredicted in absolute value. This pattern suggests that the test score distributions were short-tailed, at least for the upper tail. Examination of the distributions confirmed that fact for both the test scores and high school average.[1] The lower section of Table D shows that, when the original distributions were Normalized, predictions of restricted sample statistics were reasonably accurate.

Insert Table D

It appears that neither the estimates of tail effects made by Hedges and Friedman (1993b) nor estimates produced by the proposed model are likely to be robust to violations of the Normality assumption--at least not at the extremes of the distribution. In actual practice a given score distribution may be either short-tailed or long-tailed, and the effects of sample restriction would need to be modified accordingly. Nonetheless, it is certainly pertinent to ask what general effects sample restriction might be expected to have, distributional idiosyncracy aside. It is first desirable to check the internal validity of the model; that is, to what degree are estimates of the effects of sample restriction accurately reproduced within a data set known to be normal.

Reproducibility. If one is thinking of "restricted sample" simply as a subset of a larger population of examinees, then any variation in the observed gender differences for the restricted and unrestricted group would, indeed, be attributable to sample restriction alone. The only question would be how accurately one could reproduce changes in indicators of gender difference from knowledge of the process whereby the sample was restricted. Such accuracy can be ascertained by simulating a sample restriction.

Table E shows the results of three such simulated sample restrictions, using the aforementioned NELS data base with HSA and the four test variables Normalized. Results of the first are shown in the column headed "NELS/HSA." It involved the explicit selection of the top 10% on an equally-weighted composite consisting of HSA and the average NELS test score (also equally weighted). The model predicted the values of D_r shown in the first of the two paired numbers in each column. The actual computed value in the sample so restricted is shown in parentheses. Note first that the predictions were fairly accurate, considering that the effects of sample restriction were substantial in this simulation involving of a fairly extreme selection ratio (P = .10).

Insert Table E

Note also the compensatory effects of sample restriction: the mean difference becomes larger, favoring women, on HSA where women tended to be higher, and becomes smaller, favoring men, on test scores where men tended to be higher. This effect seems somewhat counter-intuitive. One might imagine that selecting the women and men who are outstanding on relevant proficiencies would make them more similar in the selected group. Instead, they have become even more different on the two measures where there was some difference originally. By adjusting the selection procedure, one can reduce the difference on one measure; the trade-off is a larger difference on the other measure, as illustrated in the next simulation.

The second simulation differs in two respects. First, it involves a more limited restriction: half of the original group was selected. The effects on D are less severe, and they are predicted with considerable accuracy. Another difference is that, in this case, sample restriction is based on only one measure, HSA. Nonetheless, the effects of implicit selection are apparent in an altered pattern of gender difference on the test variables. Having selection weight on HSA has reduced the gender difference on that measure, but the difference on the test measures is typically larger than in the original group. The implicit effects on the test variables appear to be predicted with essentially the same accuracy as the effect on HSA, the variable actually used in the sample selection.

The final simulation adds a different dimension. In this case the restricted sample was that group of survey respondents who reported having taken either the *American College Test* or the *Scholastic Aptitude Test*. Here again, the D_s in the restricted sample were quite accurately predicted, even though the actual selection process had no direct connection with this survey or these measures. As we noted, all of the results in Table E were based upon Normalized data. In order to check again on the importance of Normality, the last of the three simulations was repeated with the original, non-Normalized data. With Normalized data, the average absolute error in predicting D was .003; with non-Normalized data, the average absolute prediction error was .005. The accuracy of the predictions suggests that the model may be robust to violations of Normality with samples restricted at this level of selectivity (P =.60).

Overall, these results indicate that the effects of sample restriction on the pattern of gender differences in a set of measures can be predicted with considerable accuracy a) if the measures are Normally distributed or selection is not severe and b) if the relationship of the measures to the selection process (r_{XY})

is known. Thus the model appears to provide a useful basis for evaluating the effects of sample restriction *alone* on the pattern of subgroup differences; that is, independent of other possible influences. We move now to a more systematic examination of the nature of those effects.

Effects of Sample Restriction

In evaluating the possible effects of sample restriction, it is useful first to recall the situation of interest. A restricted sample may result from explicit selection on a known variable, or it may come about through a complex selection process, only partly understood. The restricted sample may show either higher or lower average achievement than the original sample. We focus here mainly on samples from the positive end of the achievement scale, though the principles are the same at either end.

The object is to estimate what effects are likely to be observed regarding gender differences on Y, a variable that has some direct or indirect relationship to the restriction process (i.e., some weight in the selection decisions). Since many variables with quite different patterns of gender difference may be involved in sample restriction, the ratio of females to males (F/M) in the restricted sample is not set. It may vary widely, and in fact, is a critical aspect of gender difference and similarity in the restricted sample. Finally, there are alternate standards to consider in evaluating any observed mean difference.

In estimating likely effects of sample restriction, there are two broad questions of interest. The first is what changes in the pattern of gender difference and similarity might one expect from sample restriction, per se, independent of other possible influences? That is, what changes would one expect due only to changing analytic relationships among means and standard deviations, assuming no other influences are at work such as variations in the test construct, changes in the selection process, departures from Normality, or differential learning of women and men over time. The second broad question is what effects of sample restriction can one likely expect in the types of sample restriction observed in actual situations, where other influences do come into play? We address these two questions in turn.

Analytic Relationships

As outlined in Table C, there are several input and output variables of special interest. The three key output variables are the F/M ratio, D_r, and D' in the restricted group. The latter two indicators of standard mean difference, D_r

and D', differ only in their use of different standards; that is, whether they are based upon standard deviations in the restricted versus the original groups. The difference between D_r and D' indicates the effects of range restriction.

There are five important input variables. The first four are: D_o in the original group; SDR_o in the original group; P, the proportion selected; and r, which represents the weight that Y had in the restriction process. Both P and r may differ for females and males. In the interest of simplifying, and because it seems likely to be less consequential in most situations, we have not attempted to illustrate the effects of differences between r_F and r_M (though the empirical simulations reported here do include any such variation). In some situations differences between r_F and r_M may be important; for example, where selective entry into a graduate field is based on different considerations for women and men. In practice, such differences in r_F and r_M are not likely to be easily estimated.

Gender difference in P is another matter. Difference in P for women and men directly determines F/M, the representation of women and men in the restricted group. As we shall see, the effects can be quite consequential. So this difference between P_F and P_M is most easily expressed in terms of its resulting effect on F/M, the fifth important input variable. As we noted, it is an unusual characteristic of the situation that this index of gender balance may be either an input or an output measure.

The equations previously derived were used to describe the relationships of interest between these five input and three output measures in both tabular and graphical form. Table F, for example, shows values of D_r and D' in the restricted sample for various combinations of D_o, SDR, P, and r_{XY}. Different values of F/M in the restricted sample are represented in Tables F, G, and H. These tables testify to the complexity of the effects of sample restriction on gender differences.

Insert Tables F, G, & H

Fortunately, these complex effects can be represented, in simplified form, as three additive components that match quite well with an intuitive interpretation of what happens to the observed gender difference in a restricted sample. Furthermore, the components are expressed in terms of the five key input variables identified earlier. Thus, the change in standard mean difference from the original to the restricted group is given approximately by

$$D_r - D_o \doteq A \cdot D_o + B \cdot \ln(SDR) + C \cdot \ln(P_F/P_M) \qquad (5)$$

where A, B, and C are functions of P and r, which define the nature of the sample restriction.[2]

The first term on the right side of (5) represents the *direct effect of sample restriction* which is proportional to D_o, the mean gender difference observed in the original group. The second term represents the *indirect effect of differential variability* expressed as *SDR*, the ratio of female to male *SD*s in the original group. The third component is associated with the differential selection rates, P_F and P_M, for females and males. To better appreciate the character of this third component, it is useful to rewrite (5) as

$$D_r - D_o \doteq A \cdot D_o + B \cdot \ln(SDR) + C \cdot \ln(F/M_r) \qquad (6)$$

The third component is here represented in terms of F/M_r, the resulting gender balance in the restricted group--an outcome measure. It is this outcome measure that has a critical trade-off relationship with D_r. For that reason, we refer to this third component of sample restriction as the *reciprocal effect of gender balance*. Together the three components explain any change that occurs in D due to sample restriction alone. Each of these three deserves some further comment as to how it might be expected to work in practice.

Direct effect of sample restriction. Assume there is some mean difference, D_o, on a Normally distributed variable in a general population that is half female and half male. How would one expect that D to change on a comparable measure in a restricted sample, also half female and half male? That is, what are the likely statistical effects of restriction, independent of other possible effects such as differential learning, a change in the test, and so on?

For the present we assume that there are equal numbers of women and men in the restricted group because there is no reason, a priori, to assume otherwise. For example, the fact that the mean score for women is typically somewhat lower than that of men on the mathematics sections of college admissions tests does not mean that there are likely to be fewer women in the sample of students who go to college. In fact, more women than men attend college because many other variables come into play. Figure G illustrates the direct effects of sample restriction--the first component of Equation (6). We start with the assumption that both F/M and *SDR* are equal to 1.00, which sets the second and third terms of (6) equal to zero.

Insert Figure G

Under these conditions we see in Figure G that sample restriction, per se, always works to increase the absolute size of *D*. The amount of the increase depends upon two factors. The more extreme the selection, the larger the absolute increase in *D*. Also, the higher the value of *r*--the weight of *Y* in selection--the larger the increase in *D*. These two effects interact because if a variable has no weight in selection, in effect, the sample is not restricted with regard to that variable. Thus, *A* in Equation (6) represents the joint action of *r* and *P*. Finally, these effects can be expressed as a percentage change for any value of D_o . As indicated in Equation (6), the amount of change is always proportional to D_o .

The general impression of Figure G is that the direct effects of sample restriction are highly dependent upon *r*. If a variable has a moderately strong weight in the process whereby the sample was restricted, the observed mean gender difference on that variable is likely to undergo a substantial proportional increase.

Indirect effect of differential variability. The effects of differential variability are indirect in the sense that changes in *D* are associated with a secondary aspect of gender difference in the original group. Figure H illustrates the nature of this effect, the second component in Equation (6). It shows the amount of change in *D* (that is, $D_r - D_o$) that can be expected with different *SD* ratios. This particular illustration describes the case where $D_o = .00$ and $F/M_r = 1.00$, thus setting the first and third term of (6) equal to zero. With an appropriate constant added to the ordinal scale, the figure can apply generally since the effect of *SDR* is constant for all values of D_o and *F/M*.

Insert Figure H

Since *SDR* appears normally to range from about .85 to 1.00 (Willingham & Cole, In Prep.), its effect will normally be to move *D* in a negative direction to the extent that it is lower than 1.00. That is, gender differences favoring males would become larger; gender differences favoring females would become smaller. Figure H shows how *SDR* interacts with *P* and *r*, its effect on *D* being enhanced for small *P* and large *r*. In terms of Equation (6), *B* is larger when *P* is smaller and *r* is larger.

Overall, Figure H suggests that differential variability is not likely to have a large effect when sample restriction is moderate (say, $P=.50$). But when the restricted sample represents a small proportion of the original group (e.g., $P=.10$), differential variability can have a substantial effect. For example, if a measure is

correlated .70 with the selection variable, F/M_r = 1.00, and SDR = .85, selection of the top 10% would result in a standard gender difference of -.26 even though D_o was zero in the original group.

Reciprocal effect of gender balance. There are two natural ways to think about and describe gender difference and similarity in a restricted group. One is to compare the difference in the observed means of women and men. Another is to compare representation of women and men. Each has its rationale as a relevant measure of interest. It is hard to argue as a general principle that it is preferable to have equal representation or equal means in a restricted group. As illustrated earlier in Figure E, these are reciprocal measures. Given a gender difference in an original group, Equation (6) illustrates that either one of these measures can be equalized in a restricted group--but only at the expense of the other. A choice between the two depends on the situation and the consequences.

As we have noted, F/M in the restricted group is directly determined by P_F and P_M. In real situations involving restricted samples, there are many reasons why those proportions may vary. Those reasons include individual choices, social constraints, institutional decisions, and so on. Figure I shows the resulting trade-off between D_r and F/M, which are plotted on the two coordinates. Panels A and B show the relationship between the two outcome measures when there is a moderate (.50) and a high (.90) level of r. Similar relationships obtain for positive values of D_o. We assume here no difference in variability in the original group and a moderate selection rate of .50.

Insert Figure I

Look first at the bottom line in Panel B which represents an original D of -.40. If selection operates so as to yield a restricted sample with equal representation of women and men, that equalization comes at the expense of a larger D_r of -.57. A lower F/M ratio of .84 would maintain D_r at the original level of -.40.

Other conditions constant, it is apparent from Figure I and from Equation (6) that the trade-off between D and F/M is the same, regardless of the original level of D_O. The important distinction illustrated in the two panels is the fact that the stakes are high when the correlation is high, but there is a less consequential trade-off when the correlation is lower. In other words, if the measure in question has little weight in the restriction process, there is relatively little variation in D_r as the representation of women and men varies. In the extreme case--for a measure uncorrelated with selection--any mean difference in

the original group should be manifested at the same level in the restricted group, regardless of the number of females and males represented.

The reciprocity of these two outcomes and the enhanced stakes associated with the weight a measure has in sample restriction both invite serious consideration as to the policy implications of selection procedures. Most obvious would seem to be the tough decisions that arise when there is an imbalance in women and men who have reached a given level of proficiency. In some cases, it might be important to have equal representation; in other cases it may be more important for both women and men to be at some desired level of proficiency. In this situation the importance of the proficiency is a critical consideration.

Another policy implication concerns the consequence of heavy weight on one or two measures. These relationships indicate that the more narrowly selection decisions turn on one or two relevant skills, the more difficult it will be to balance both representation of women and men and their average proficiency on the chosen skills. We have another aspect of the outcome measures to consider; namely, how to interpret the observed standard mean difference in light of the fact that the "standard" has shifted.

Range restriction. It is not unreasonable to suspect that range restriction may often play an important role in apparent increases in standard mean difference in selected groups. It is clear that one regular outcome of sample restriction will be reduced standard deviations. If a new, reduced standard is used in judging a mean difference, a larger D_r will necessarily result. Judging a mean difference in a restricted group with the old standard has a different rationale, as previously noted, and the alternate statistic, D', avoids the effect of range restriction.

It is not so much a question of which metric is correct, but what one may learn from looking at both. It is clear that they are different. Examining Tables F, G, and H, where the two sit side by side, gives an impression of much smaller gender difference in selected groups when the standard is the same as that used for unrestricted groups. That is particularly the case when the selection ratio is low and when the measure in question has a heavy weight in sample restriction.

That result is well illustrated in Figure G where *all* of the effects shown are due to range restriction. Range restriction is particularly, though not uniquely, associated with reducing the sample to one tail. Restriction interacts with differential variability in its effect on D_r, and sometimes the two work in opposite directions. Thus, range restriction is not a component that can be expressed as some given percentage of the overall effects of sample restriction. Two examples give some feel for the likely contribution of range restriction.

Assume that a high school senior class, equally represented with women and men, has taken a natural science test with these results: D = -.20 and SDR = .90. Assume also that half of both the female and the male seniors go to college. What effect would one expect on D_r, the standard mean difference for the science scores among those freshmen? One can reasonably assume that science proficiency, as represented by the test, had only a modest relationship with the decision to attend college; say, an r of .50. Table G indicates that the effects of sample restriction would likely increase the mean difference six points to -.26, two of those points or about one-third being due to range restriction.

Changing the situation, assume that the same class has the opportunity at the beginning of the senior year to take an Advanced Placement course in Chemistry. Knowing this to be a very difficult course, only 10% of the seniors decide to sign up--those who are interested in science and have done well in the subject in the past. Among those who do, assume there are about four women for every five men. What effect might one expect on D_r, the standard mean difference for the science score among the students in the course? In this situation it is reasonable to assume that high performance on such a science test would have a fairly high correlation with the decision to take the course; say an r of .90. Table F indicates that the effects of sample restriction would change the mean difference 27 points from -.20 to -.47. Twenty of those points, or about three-quarters, are due to range restriction. From this perspective, the difference in mean proficiency level between the women and men entering the course is not nearly as large as would be suggested by routinely computing D with a denominator based on the smaller SDs of the restricted group.

Needless to say, somewhat different, equally plausible assumptions would make some difference in these estimates, but would not alter the basic point. Range restriction can be a minor factor in some situations, a major factor in others. Where appropriate and possible, it is useful to consider both standards, restricted and unrestricted, in evaluating an observed mean difference.

Predictions With Actual Data

Is it possible to see in actual data evidence of the analytic effects previously described? To what extent do such effects appear to account for variations one observes in the pattern of gender differences in restricted versus nationally representative samples? In order to address these questions, we need to meet two general requirements.

With actual data, there are several reasons why the pattern of gender differences in restricted samples might differ from that normally observed in tests

administered to nationally representative groups of students. Other than sample differences, the main possibilities are cohort differences and construct differences. The first requirement, then, is to identify comparison data in which those two factors are controlled insofar as possible. The second requirement is to find a plausible basis for estimating the parameters used in the model in order to predict the effects of sample restriction.

The obvious place to start looking is near the end of secondary school where there is much data available on representative as well as restricted samples. In the data base previously cited (Willingham & Cole, In Prep.), there were 74 tests in 15 categories that had been administered to representative samples of "12th grade" students (actually some were 11th graders). Self-selected samples of those 12th graders took a variety of more difficult but generally similar tests for college admissions purposes.

Many of these are achievement tests from the Advanced Placement Program (AP) or the Admissions Testing Program (ATP) of the College Board. They are taken by widely varying numbers of very able students. They cover a variety of subject areas such as Chemistry, Spanish Literature, and Calculus. These tests are poor choices for the task at hand. One reason is that the constructs do not match the tests administered to 12th graders generally. Also, we have, at best, a very shaky basis for estimating the necessary parameters--mainly because we have little knowledge of the process through which individual students decide to take these tests.

For example, it is reasonable to assume that individual students elect to take AP tests partly on the basis of whether the course is offered in their school, whether they have done well in similar courses, whether they are interested in the subject, whether it is readily scheduled, whether they intend to seek course credit in college, and so on. These factors would likely have quite different effects on *r*. With little hard information on how such matters work in actual practice, estimating *r* (i.e., how strongly competence in the subject is related to membership in the test-taking group) is largely a blind guess. The ATP subject tests pose similar problems. These various sources of uncertainty in estimating *r*, a critical parameter, illustrate why predicting the effects of sample restriction on a particular test is likely to be problematic.

A more promising possibility is to estimate restriction effects for that group of tests taken by college-going 12th graders for admissions purposes; namely ACT, PSAT, and SAT. We had data on 11 such tests taken by the broad college-going group rather than some more selective subsample such as those who take the AP Chemistry examination. High school average is also available for both the restricted and unrestricted group of 12th graders. The objective is to predict,

through the following procedure, what gender difference, D_r, would be expected for each of these measures on the basis of sample restriction alone.

First, each of the 11 tests was matched as closely as possible with one of the 15 categories of tests administered to representative 12th graders. This yielded the pattern of categories and tests shown in Table I. High school average was added to the analysis.

Insert Table I

The number of non-selective tests, and associated data sets, in each 12th grade category is indicated in parentheses--all together, 47 tests and two very large sets of data for HSA. The next step was to estimate the likely range of D_r values in each category if all measures in the left column above were subjected to sample restriction typical of college admissions test-takers nationally. This involved predicting a value of D_r for each of the 49 data sets using, as the first two parameters, D and SDR in each of the original representative samples.

The other three necessary parameters are P, F/M_r, and r. Since the same restricted sample of admissions test-takers applies in all cases, a common set of the first two parameters was used in all 49 predictions. A common r parameter was used within each category. Two methods were used to estimate these three parameters. In Method A we estimated P and F/M_r from ACT and College Board program statistics, and r from simulations assuming that students decide to take an admissions test on the basis of their HSA, previous test performance, and chance factors.[3] In Method B we estimated P, F/M_r, and r on the basis of those students in the 1992 NELS sample who reported having taken either the ACT or the SAT.[4] Again, the rationale of this procedure is to estimate what standard mean difference, D_r, one would expect in restricted samples for the tests administered to representative samples, and then to compare the predicted values with those actually observed in similar tests.

Figure J provides several types of information to summarize the results. The 13 stars (★) on the various category lines show the *actual* values of D_r in the restricted samples. For each test category and high school average, the symbol ⊗ represents the average standard difference, D_o, for all measures in that category based on representative samples in Grade 12. The solid lines and dashed lines represent, for Methods A and B respectively, the mean ± 1 SD for all *predicted* values of D_r within each category based on estimates of the effect of sample restriction.

Insert Figure J

First a comment on the two methods of estimating parameters. Results for predictions based on Method A and Method B (solid and dashed lines, respectively) were quite similar as to direction and range; Method A typically predicted somewhat larger effects of sample restriction than did Method B. Both methods have weaknesses. Method A requires an educated guess as to how much weight HSA, test scores, and other unknown factors including chance, have on the decision to take an admission test. Method B is based upon 69% of the NELS sample--those who had complete test, grade and questionnaire data. This group may or may not represent well high school seniors nationally. The accuracy of Method B also depends upon correct reporting of test-taking by students. The most accurate predictions probably lie somewhere between these two methods.[5] Figure J suggests four main results:

- In all seven categories, the model predicts that D will move slightly or moderately in a negative direction, favoring males.
- As predicted, in all seven categories the observed D_rs are more negative than the average D_o for representative samples at Grade 12.
- In most categories, observed values of D_r for tests administered to restricted samples lie within the range of predicted values.
- For two types of measures--language use and high school average--the observed values of D_r tended to be noticeably more negative than would be predicted only on the basis of sample restriction.

We have several observations about these findings. The predicted trend to more negative values of D_r is apparently due mainly to the somewhat greater variability of male scores in most of the categories and to somewhat greater representation of women among test-takers (F/M is about 1.16 in test program data). In this illustration the effect of restriction, per se, is not likely to be large since the selection ratio is not small (60% of students in the NELS sample reported taking an admissions test). The fairly wide range of such predicted values underscores the fact that tests within categories vary in character, and that prediction of standard mean differences in restricted samples on the basis of sample restriction alone is an unreasonable expectation. Nonetheless, the findings do show reasonably good consistency overall between trends predicted by the model and trends widely observed in gender differences on selective tests.

The two inconsistencies, language use and high school average, are curious. The language use tests taken by selected samples do not appear to be qualitatively different from those regularly administered to general populations of high school seniors. There is some suggestion of a similar result for other types of language tests. The reading subscore of the SAT showed a *D* of -.06 in 1992, placing it and the ACT reading test at the low end of the predicted range. There are no data on essay writing for a college-going sample comparable to that on which these data are based, though what data are available show markedly lower *Ds* in selected samples than are typical in general populations. It may be that such tests tend to have short-tailed score distributions at the top. Reading and language use tests do not always discriminate well at the top of the scale (Donlon, 1984; Lord & Wild, 1985). On the other hand, it is possible that the construct is somewhat different when such tests are designed for more able students.

High school average is a similar but even more curious case because it is, in fact, essentially the same measure for the unrestricted and restricted samples. NELS and HSB data show *Ds* of .31 and .35 for HSA based on high school seniors; ACT and College Board program data show *Ds* of .15 and .16 based on college-going students. There may also be a short-tail problem with HSA due to a ceiling effect, but some of our analyses seem inconsistent with that as the explanation. Another possibility is that males may be more prone to exaggerate when self-reporting their HSA in the college admission context. If so, that would tend to wash out some of the positive *D* observed in survey data. Available data appear to be inconsistent on this point.[6]

Summary

Our purpose has been to work toward an improved understanding of the effects on observed gender differences of sample restriction as it may operate under a variety of conditions. This requires taking into account three important features: effects of implicit selection on new measures in new situations, the effects of gender balance on the outcome of sample restriction, and the standard one chooses to employ in evaluating mean differences. Building on recent work on differential variability and explicit selection in the tail of a distribution (Cleary, 1992; Feingold, 1992b; Hedges & Friedman, 1993b), we proposed a sample restriction model that incorporates these features. The model includes five input variables and three output variables.

In simulated sample restrictions with test and grade data, the model proved to be quite accurate in reproducing standard mean differences and other statistics in the restricted group. This was true even in the case of implicit effects on variables not actually involved in the restriction process. These findings support the internal validity of the model. Also, the model appears to be relatively robust

to violations of Normality assumptions with selected proportions in the range of .50, but much more sensitive if the selected proportion is small (i.e., .10 or less).

Strictly speaking, the statistical effects of sample restriction apply in a predictable manner for a given measure, within a particular sample, wherein females and males are selected in a similar manner. Additional contingencies arise when such a model is used to make external predictions; for example, in attempting to account for differences in observed *D* for two similar tests administered to unrestricted and restricted samples. In such situations a key input statistic is apparently *r*, the correlation between the original variable of interest and the restriction process. This parameter has consequential effects, is typically unknown, and is difficult to estimate. It is also apparent that external predictions can be in error because of subtle differences in the selection process for women and men or in the constructs represented at the input and output ends of the prediction of interest.

Because of these considerations, it seems clear that the proposed model is more useful for understanding the general principles that are likely to influence gender differences under sample restriction than for predicting a restriction effect in a particular set of test data. Nevertheless, our results show reasonably good consistency overall between trends predicted by the model and trends widely observed in gender differences on selective tests.

The model suggests several principles by which sample restriction affects gender differences. It was possible to demonstrate, analytically and graphically, that there are three major components to such effects. First is the direct effect of restriction, per se. When the same proportion of women and men is maintained from an unrestricted to a restricted sample and there is no difference in variability, restriction always increases the absolute value of the apparent mean gender difference. All of that increase is due to range restriction, not to any change in the relative performance level of women and men. It does not necessarily follow, of course, that the character and the consequences of that difference are the same at different levels of proficiency.

The second major component is the indirect effect of differential variability in the original unrestricted sample. The typical effect of differential variability is a negative shift (i.e., favoring males) in the standard mean difference in the restricted group. Either of these two components or types of effects may be small or substantial, depending upon the circumstances. The two components may have a cumulative effect or they may counteract each other. Consequently, it is not possible to formulate any general rule as to the extent to which changes in the pattern of gender difference are due to one factor or the other.

The third component of the effects of sample restriction on mean gender difference is related to the relative proportion of women and men who are selected. If more men are selected ($P_M > P_F$), then D_r moves in a positive direction (i.e., favoring females); and vice versa. This third effect of sample restriction is quite different from the first two because it is not clear that it is always desirable to minimize the mean difference, D_r.

Gender difference in any restricted group can be evaluated on the basis of two outcome measures: the standard mean difference, D_r, and the representation of women and men who are selected, F/M_r. The characteristics of the model make clear that these are reciprocal features of gender difference and similarity that can vary substantially, depending upon the nature of the sample restriction. If a sample is selected in such a manner that one of the outcomes tends to favor one gender, the other outcome will necessarily move in the opposite direction *if* other factors remain constant.

It is important to understand, however, that the reciprocal relationship between F/M and D holds strictly only in a static situation; that is, for a given measure within a given sample. If the sample or the situation changes, D may well be affected by F/M and vice-versa, but their relationship is not necessarily determinative. Either gender may improve in relative performance on either or both measures. A good illustration of such an outcome can be observed in test results of the Advanced Placement Program over the past decade. A significant increase in representation of women (F/M) in natural science AP exams in recent years has not been accompanied by any change in D, the relative mean score level of females and males (College Board, 1993).

Another type of reciprocity was illustrated in simulations of restriction effects involving more than one selection measure. When two measures like high school average and an admissions test score show opposite gender differences, sample restriction may increase both differences or decrease one at the expense of a larger difference on the other. One important conclusion we draw is that such inevitable trade-offs bring into focus important issues concerning policy and practice in seeking gender equity. When is differential proficiency more or less important than differential representation in a restricted group? What are the costs and benefits of gender parity on one proficiency in a restricted group if it comes at the expense of disparity in another proficiency?

REFERENCES

Baratz-Snowden, J., Pollack, J., & Rock, D. (1988, April). Quality of responses of selected items on NAEP special study student survey. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Becker, D. F., & Forsyth, R. A. (1990, April). Gender differences in academic achievement in grades 3 through 12: A longitudinal analysis. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Chipman, S. F., & Wilson, D. M. (1985). Understanding mathematics course enrollment and mathematics achievement: A synthesis of the research. In S. F. Chipman, L. R. Brush, & D. M. Wilson (Eds.), Women and mathematics: Balancing the equation (pp. 275-328). Hillsdale, NJ: Earlbaum.

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In J. Pfleiderer (Ed.), Sex equity in educational opportunity, achievement, and testing. Proceedings of the 1991 ETS Invitational Conference (pp. 51-90). Princeton, NJ: Educational Testing Service.

College Board. (1983, 1993). National summary report-Advanced Placement Program. New York: College Entrance Examination Board.

Donlon, T. F. (Ed.). (1984). The Test of Standard Written English (TSWE). In T. F. Donlon (Ed.), The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests (pp. 69-84). New York: College Entrance Examination Board.

Feingold, A. (1992a). The greater male variability controversy: Science versus politics. Review of Educational Research, 62(1), 89-90.

Feingold, A. (1992b). Sex differences in variability in intellectual abilities: A new look at an old controversy. Review of Educational Research, 62(1), 61-84.

Feingold, A. (1993). Joint effects of gender differences in central tendency and gender differences in variability. Review of Educational Research, 63(1), 106-109.

Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. American Psychologist, 50(1), 5-13.

Fetters, W. B., Stowe, P. S., & Owings, J. A. (1984, September). High School and Beyond: A national longitudinal study for the 1980's. Quality of responses of high school students to questionnaire items. Washington, D.C.: U. S. Department of Education, National Center for Education Statistics.

Freeberg, N. E. (1988). Analysis of the revised student descriptive questionnaire, phase I: Accuracy of student-reported information (CB Report No. 88-5, ETS RR-88-11). New York: College Entrance Examination Board.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Han, L., Cleary, T. A., & Rakaskietisak, S. (1992). Gender differences on achievement tests: A trend study based on nationally representative samples. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Hedges, L. V., & Friedman, L. (1993a). Computing gender difference effects in tails of distributions: The consequences of differences in tail size, effect size, and variance ratio. Review of Educational Research, 63(1), 110-112.

Hedges, L. V., & Friedman, L. (1993b). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. Review of Educational Research, 63(1),94-105.

Johnson, N. L., & Kotz, S. (1970). Distributions in statistics: Continuous univariate distribution-1. Boston, MA: Houghton-Mifflin.

Kaufman, P., Rasinski, K. A., Lee, R., & West, J. (1991, September). Quality of the responses of eighth-grade students in NELS:88 (CS 91-487). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Lord, F. M., & Wild, C. L. (1985). Contribution of verbal item types in the GRE General Test to accuracy of measurement of the verbal scores (GRE Report No. 84-6P, ETS RR-85-29). Princeton, NJ: Educational Testing Service.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University.

Martin, D. J., & Hoover, H. D. (1987). Sex differences in educational achievement: A longitudinal study. [Special issue: Sex differences in early adolescents.] Journal of Early Adolescence, 7(1), 65-83.

Maxey, E. J., & Ormsby, V. J. (1971, July). The accuracy of self-report information collected on the ACT test battery: High school grades and items of nonacademic achievement (ACT Research Report No. 45). Iowa City, IA: American College Testing Program.

McNemar, Q., & Terman, L. M. (1936). Sex differences in variational tendency. Genetic Psychology Monographs, 18(1), 1-65.

Noddings, N. (1992). Variability: A pernicious hypothesis. Review of Educational Research, 62(1), 85-88. (Commentary on Feingold, 1992).

Obler, L. K., & Fein, D. (Eds). (1988). The exceptional brain: Neuropsychology of talent and special abilities. New York: Guilford.

Singer, J. E., Westphal, M., & Niswander, K. R. (1968). Sex differences in the incidence of neonatal abnormalities and abnormal performance in early childhood. Child Development, 39(1), 103-112.

Snow, R. E., & Ennis, M. (1992, April). Correlates of high mathematical ability in a national sample of eighth graders. Paper presented at From Psychometrics to Giftedness: A Symposium in Honor of Julian C. Stanley, San Francisco, CA.

U. S. Department of Education, National Center for Education Statistics. (1992). Digest of education statistics 1992 (NCES 92-097). Washington, DC: U. S. Government Printing Office.

Willingham, W. W., & Cole, N. S. (In preparation). Gender and fair assessment. Princeton, NJ: Educational Testing Service.

TECHNICAL NOTES

1. High school average and all four NELS tests showed a moderate level of negative kurtosis (-.75 to -.95). The tests showed a slight negative skewness (typically about -.16, except for the reading test at -.37); the HSA was skewed positively to a small degree (.20). In the case of the test scores, the negative kurtosis may result from the use of a Baysian method of IRT scaling in the NELS study, which tends to pull in the tails somewhat. Apparently, high school averages tend naturally to be moderately short-tailed. When Normalizing NELS test scores, we simply replaced the original score at the P^{th} percentile with the corresponding percentage point for the Normal distribution having the same mean and standard deviation. This was done separately for male and female distributions.

2. This equation was first obtained empirically, after studying Tables F, G, and H. We noted that, for fixed values of the correlation (r) and the proportion selected (P), the values of D_r given in the tables could be approximately represented as a sum of three components, related to D_o , SDR, and P_F/P_M . The size of these components varies as a function of r and P, but they are always (approximately) proportional to D_o , $\ln(SDR)$, and $\ln(P_F/P_M)$, respectively. Note that the input variables P_F and P_M cannot be varied independently in Equation (4), since it assumes that P is fixed. In other words, if you were to hold P_M constant and vary P_F , the weights A, B and C would all change, as a result of changing $(P_F+P_M)/2 = P$.

3. P, F/M and r were estimated by Method A as follows. Based on a 1992 high school senior class of 2.5 million (U. S. Department of Education, 1992), program data from ACT and the College Board indicate that about 33% and 41% of those seniors took the ACT and SAT respectively. Considering overlap testing, we estimated that 65% of all seniors took one or both tests. The F/M ratio of test candidates for both testing programs combined was 1.16. In order to estimate r, a selection variable X was defined as an equally weighted composite of HSA, NELS test composite and a random variable with D = .00 and SDR = 1.00. Selecting the top 65% on X yielded values of r_{bis} in the range of .60 to .65 for the tests and .70 for HSA.

4. P, F/M, and r were estimated by Method B as follows. The 1992 NELS questionnaire asked seniors to report whether they had taken the ACT and/or the SAT. The group who said that they had taken both defined P = .61 and F/M = 1.19. This definition of the restricted sample yielded values of r_{bis} about .10 lower than the estimates produced by Method A.

5. Method A and Method B gave reasonably similar estimates of P and F/M as described above. The difference in the estimates of these two parameters did not appear to be a significant source of difference in the predictions. An error in r_{bis} is likely to have more influence. The Method A estimate is suspect because it involves a judgmental simulation. Method B is suspect because it is based upon a sample with significant data loss, and any student errors in reporting on the tests would tend to reduce r_{bis} with an attendant underestimation of restriction effects. Curiously, it seems that not all students are accurate in reporting such information. We know of no good data on this specific point. Though there is some indication that, in general, males tend to be less accurate in their survey responses (Kaufman, Rasinski, Lee, & West, 1991; Baratz-Snowden, Pollack, & Rock, 1988). It is on the basis of these several considerations that we suggest that the more accurate predictions lie somewhere between Method A and Method B.

6. A study by the National Center for Education Statistics (Fetters, Stowe, & Owings, 1984) reported no gender difference in the tendency to exaggerate grades earned. On the other hand, Freeberg (1988) reported that 21% of males and 16% of females overreported. Maxey and Ormsby (1971) reported similar figures: 15% for males and 12% for females.

Figure 10-A
Differences in the Upper Tail (90th Percentile) When Female (F) and Male (M) Scores Differ in Both Mean and Standard Deviation

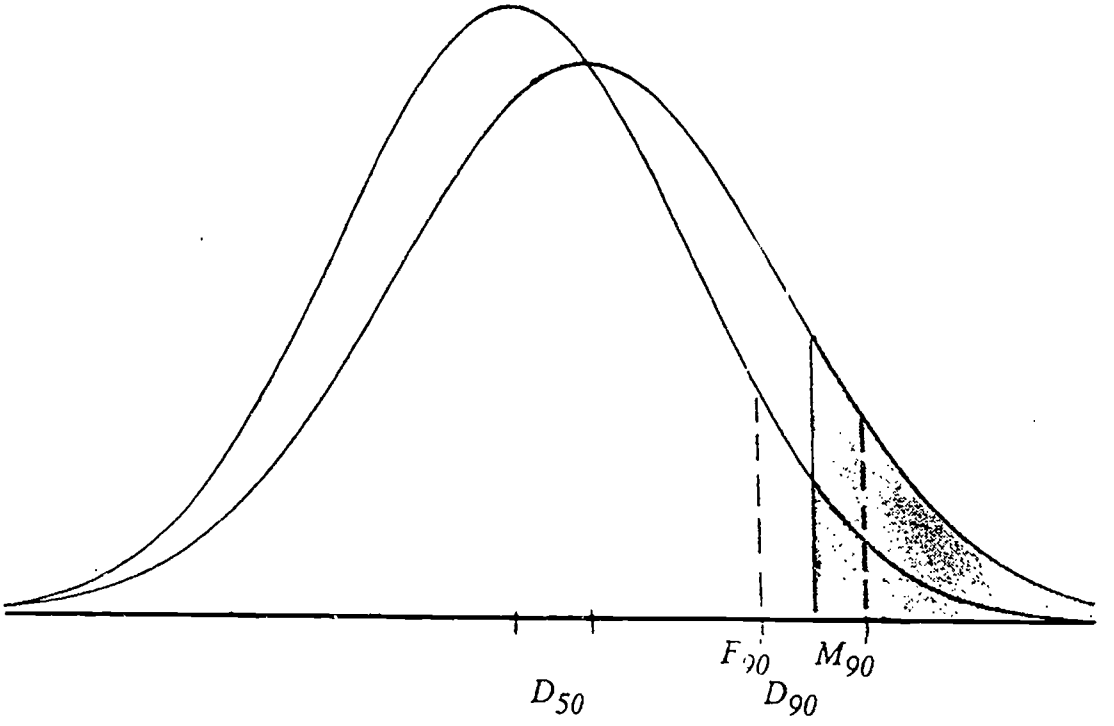


Figure 10-B
Illustration of the Largely Independent Role of Original D_o and the Standard Deviation Ratio in Determining D_r Within the Upper Tail (Top 10%)

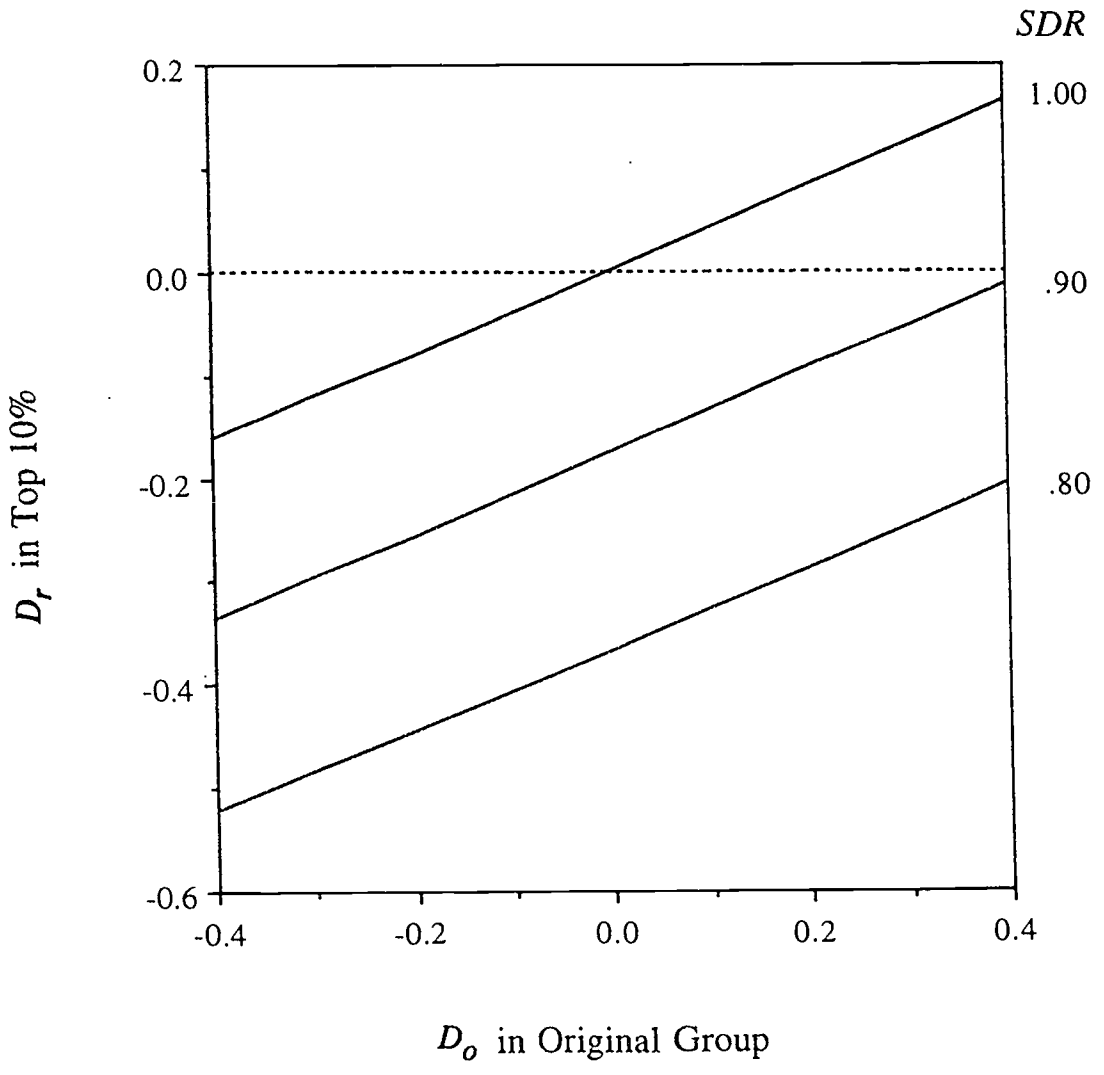


Figure 10-C
Illustration of the Largely Independent Role of Original D_o and the Standard Deviation Ratio in Determining F/M Ratio Within the Upper Tail (Top 10%)

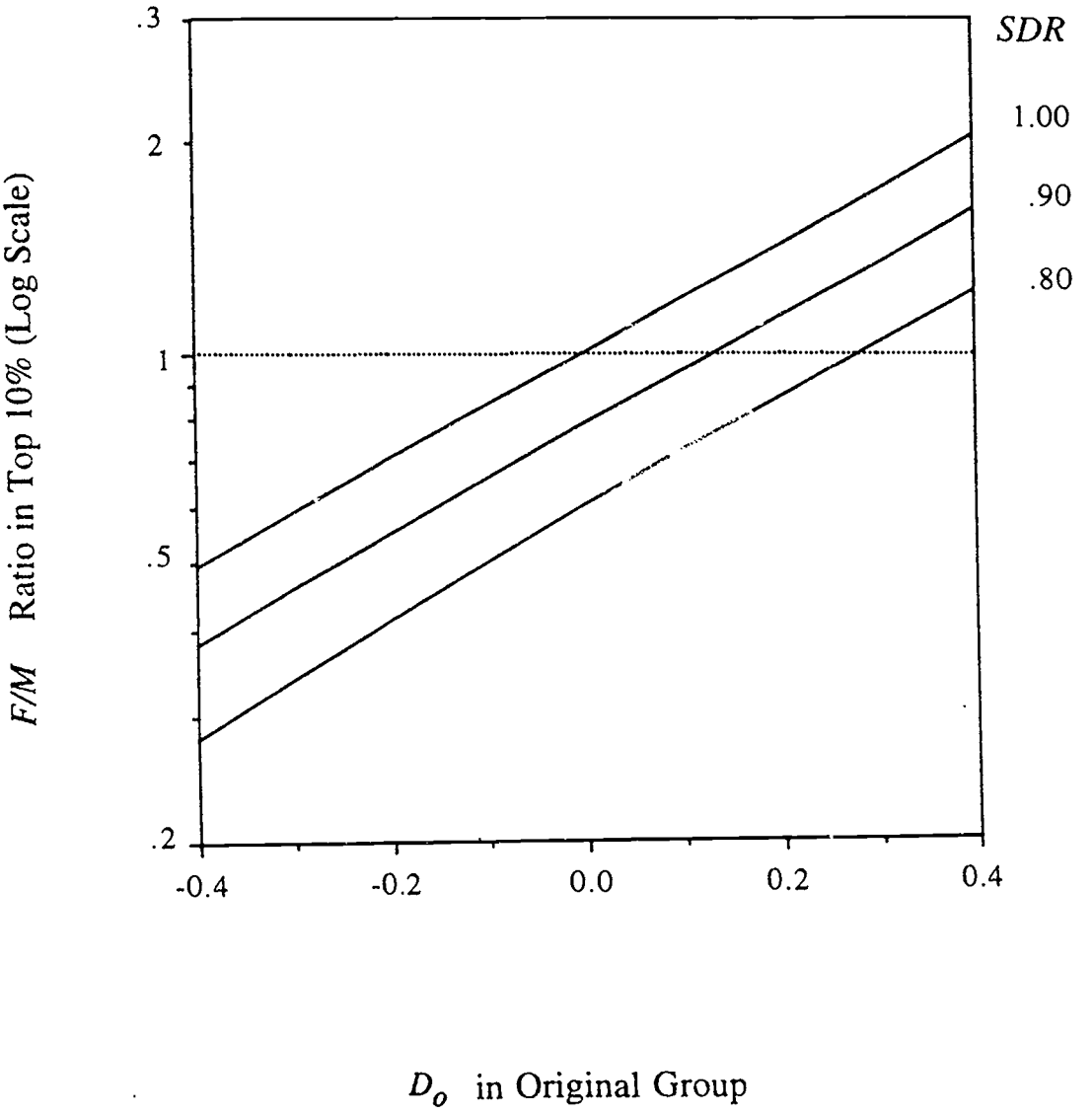


Figure 10-D
Restriction Model: Illustration of Implicit Selection on Variable Y
Due to Its Correlation With a Hypothetical Composite X,
on Which a Sample is Explicitly Selected

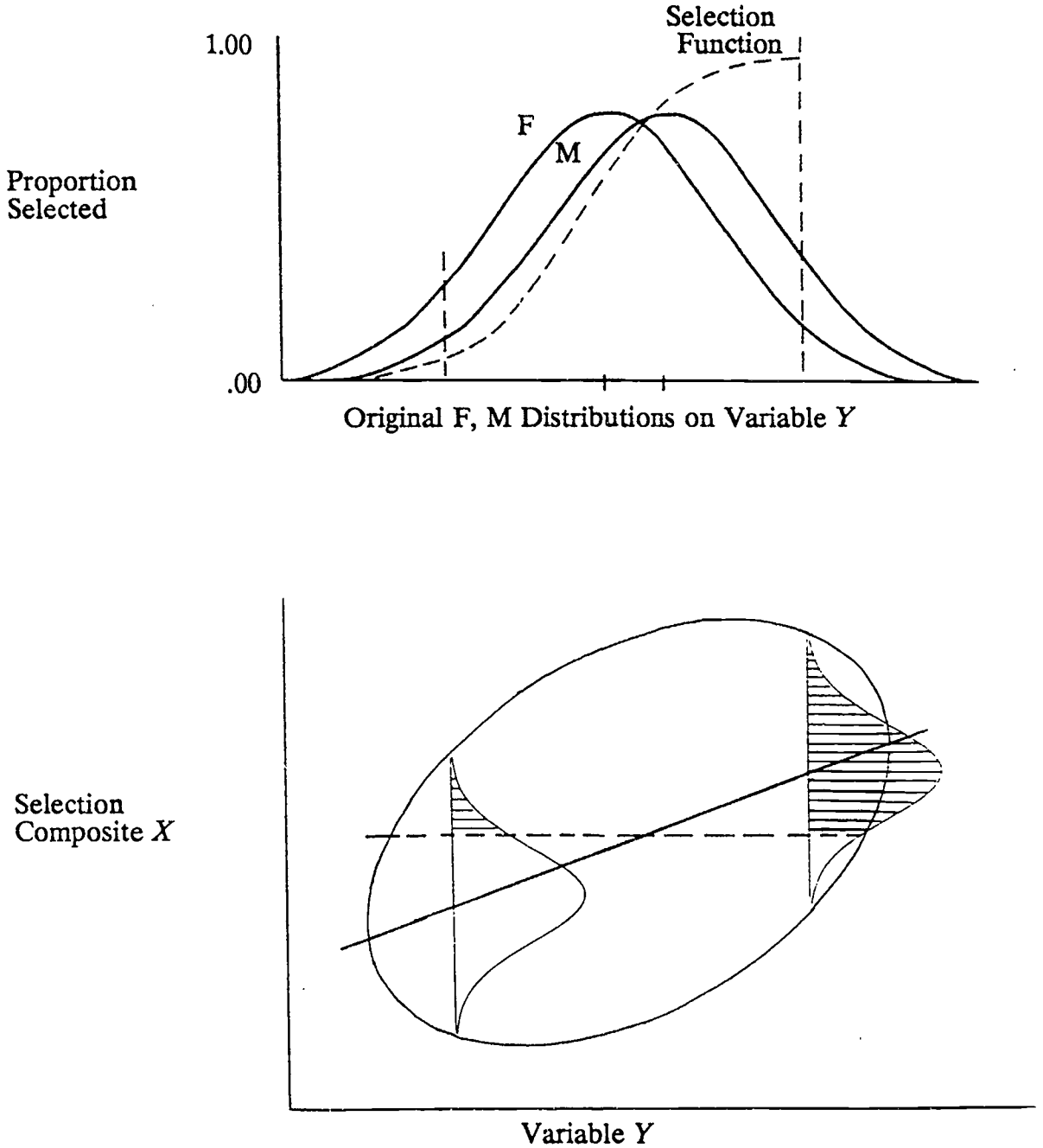


Figure 10-E
Illustration of Reciprocal Effects on Standard Mean Difference (D_r) and Female-Male Ratio (F/M_r), Depending Upon the Nature of Sample Restriction

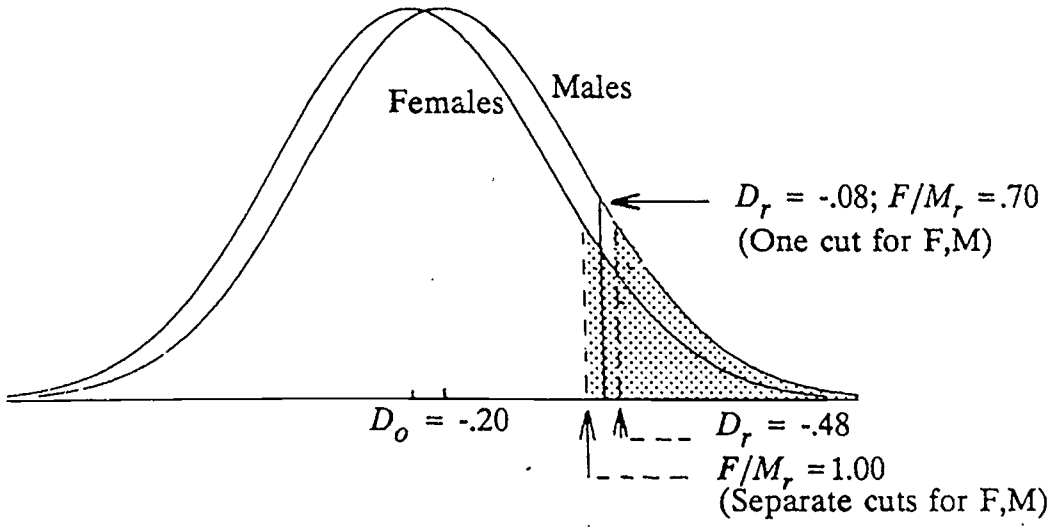


Figure 10-F
Proportion of Females and Males Taking a College Admissions Test (ACT or SAT) as a Function of (A) High School Average and (B) Composite NELS Test

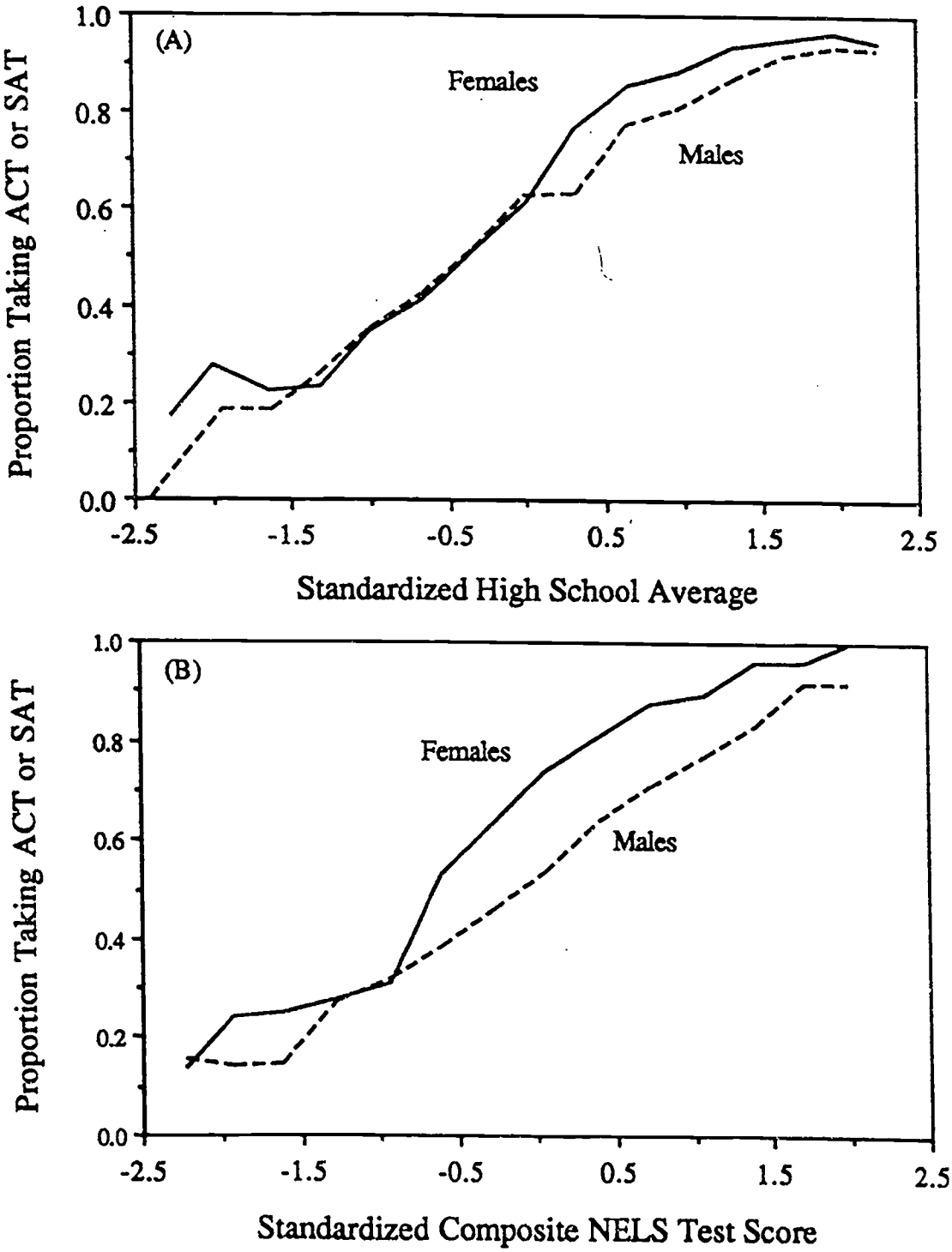


Figure 10-G
Direct Effects on Standard Mean Difference (D_r) in a Restricted Sample
Under Different Conditions of Sample Restriction
(SDR = 1.0; F/M Ratio = 1.0)

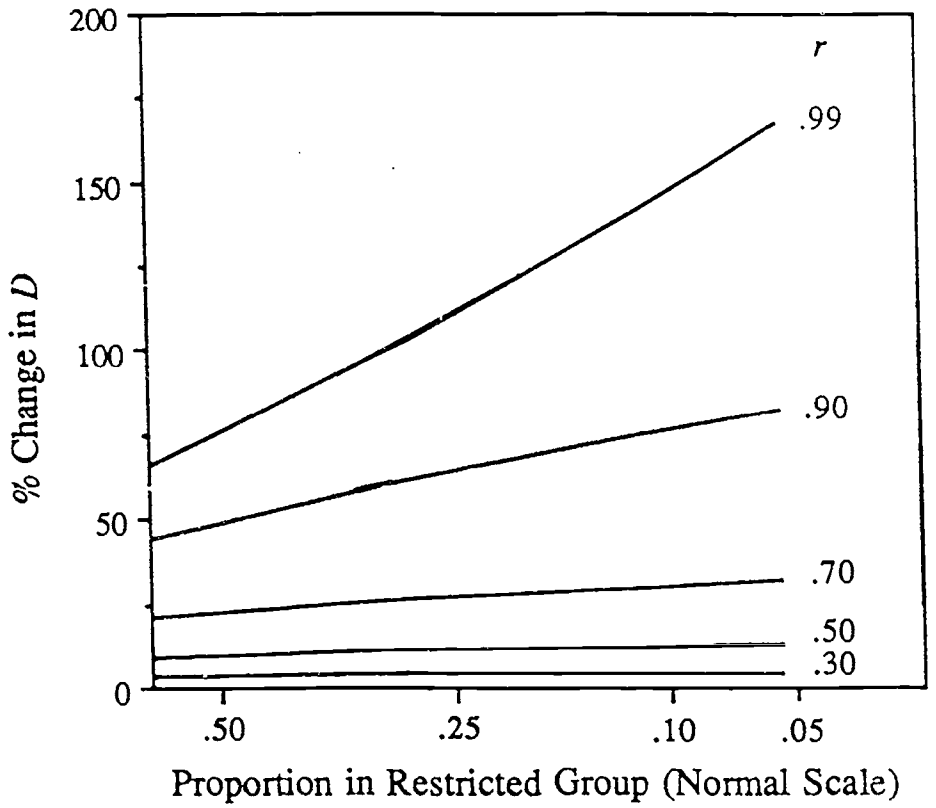
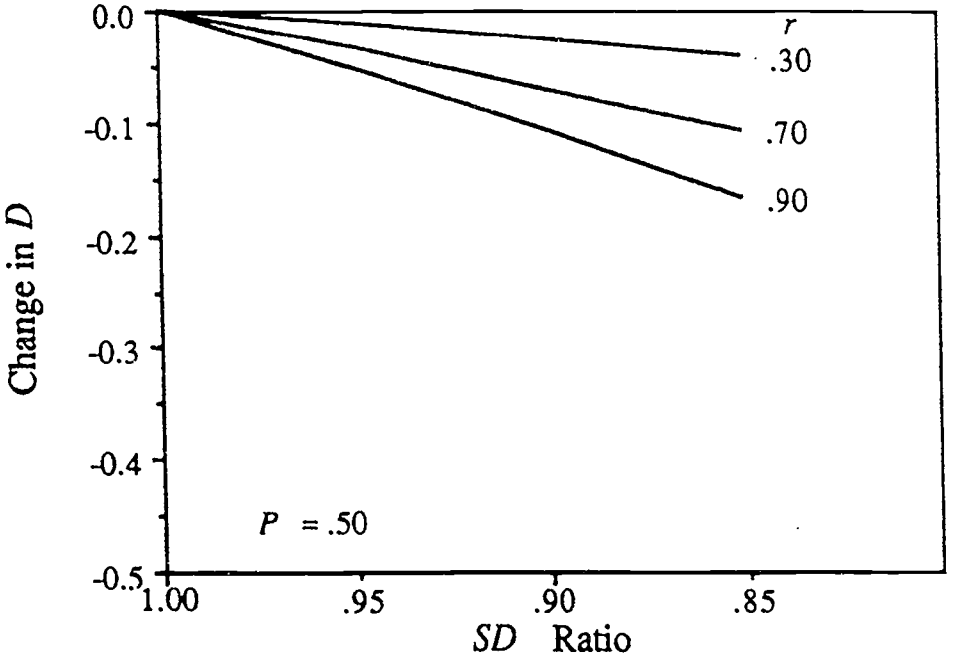


Figure 10-H
Indirect Effects on Standard Mean Difference (D_r) in a Restricted Sample Associated with Differences in Variability (*SDR*) in the Original Sample



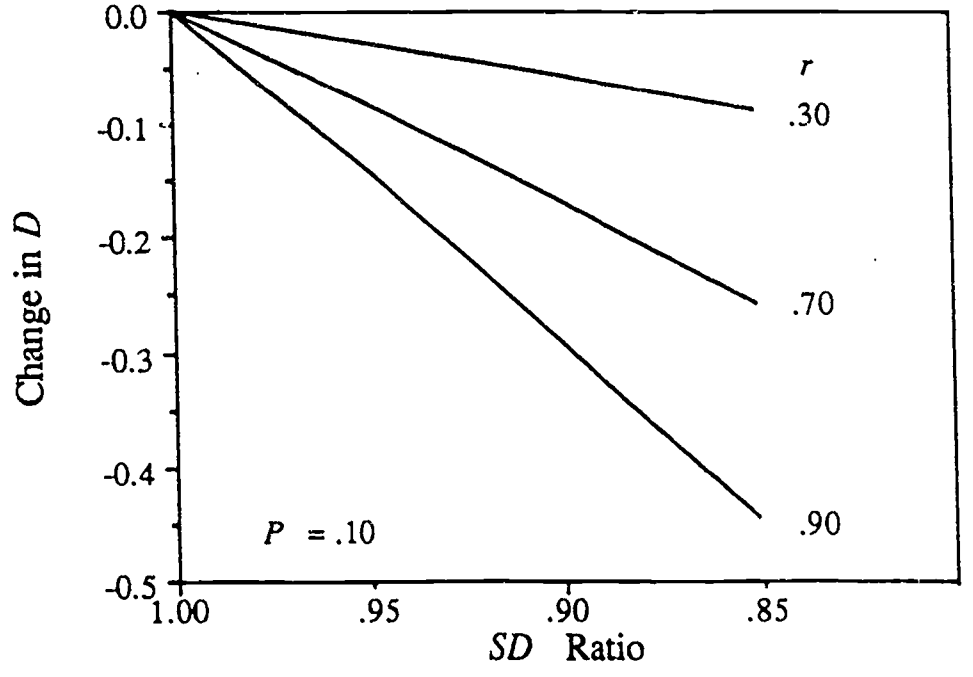


Figure 10-I
The Trade-off Between Standard Mean Difference (D_r) and Representation of Females and Males (F/M_r) in a Restricted Sample (SDR = 1.00; P = .50)

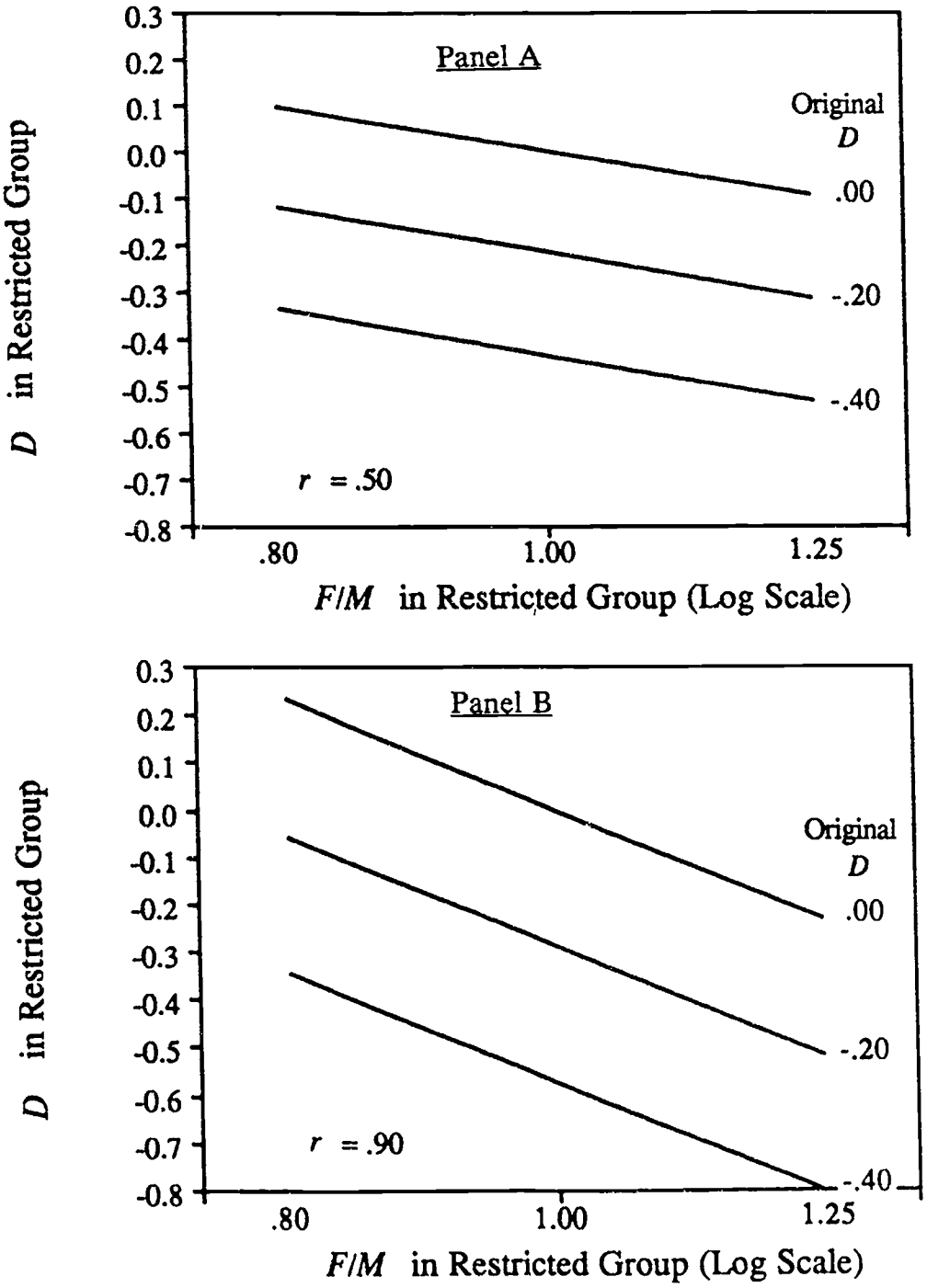
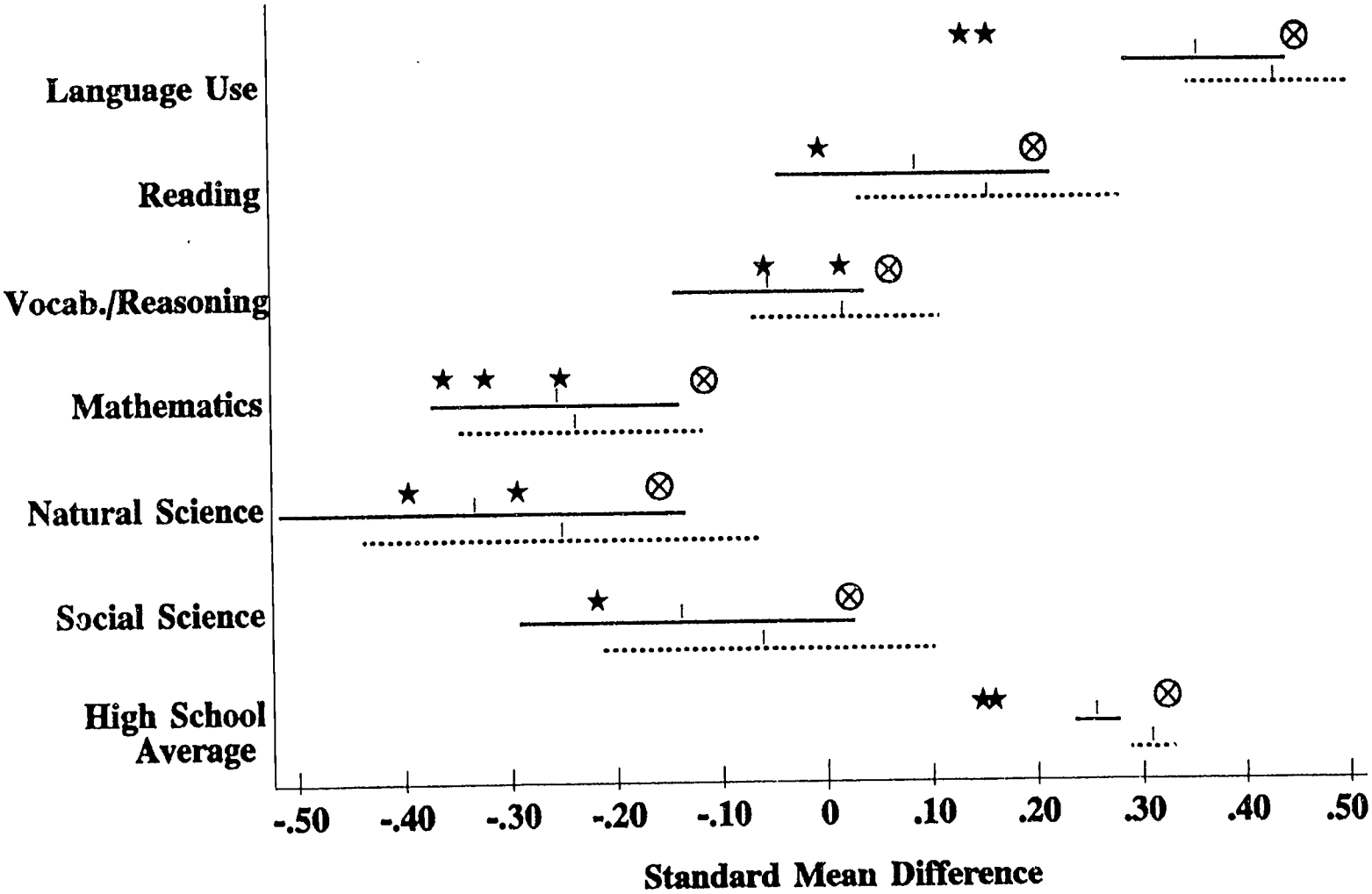


Figure 10-J
Predicted Range and Actual Mean Difference for Eleven College Admissions Tests and High School Average



*Symbols: ★ shows individual D_r for 11 tests and HSA in restricted samples;
⊗shows mean observed D_o for similar tests and HSA in representative samples for each category.
Mean ± 1 *SD* for predicted D_rs are ——|—— for Method A and .....|..... for Method B.

Table A
Standard Mean Difference (D_r) and Female-Male Ratio (F/M) in the Upper Tail (Top 10%, Top 50%) as a Function of the Standard Deviation Ratio (SDR) and D_o in the Original Group#

(Top 10%)		D_o in the Original Group:				
	SDR	-.40	-.20	.00	.20	.40
D_r	1.00	-.16	-.08	.00	.08	.16
	.95	-.25	-.17	-.09	-.00	.08
	.90	-.34	-.26	-.18	-.09	-.01
	.85	-.43	-.35	-.27	-.19	-.11
	.80	-.52	-.45	-.37	-.29	-.21
F/M	1.00	.49	.70	1.00	1.42	2.04
	.95	.43	.63	.89	1.27	1.80
	.90	.38	.55	.79	1.12	1.59
	.85	.33	.48	.69	.99	1.40
	.80	.28	.41	.60	.86	1.23
(Top 50%)						
D_r	1.00	-.24	-.12	.00	.12	.24
	.95	-.31	-.19	-.07	.05	.17
	.90	-.38	-.26	-.14	-.02	.10
	.85	-.45	-.33	-.21	-.09	.03
	.80	-.53	-.41	-.29	-.17	-.05
F/M	1.00	.73	.85	1.00	1.17	1.38
	.95	.73	.85	1.00	1.17	1.38
	.90	.73	.85	1.00	1.17	1.38
	.85	.73	.85	1.00	1.17	1.38
	.80	.72	.85	1.00	1.17	1.38

#Following Hedges and Friedman, 1993

Table B
Effect of Differential Female and Male Variability (*SDR*) on the Standard Percentile Difference (D_p) at the 90th Percentile

	D_o in the Original Group:				
SDR	-.40	-.20	.00	.20	.40
1.00	-.40	-.20	.00	.20	.40
.95	-.47	-.27	-.07	.13	.33
.90	-.53	-.33	-.13	.07	.27
.85	-.61	-.41	-.21	-.01	.19
.80	-.68	-.48	-.28	-.08	.12

Table C
Sample Restriction Model: Statistical Components

	Input Statistics, Original Sample		Output Statistics, Restricted Sample
	F, M Separately	*F, M* Difference	*F, M* Difference
1. Sample Size	N_F, N_M	F/M_o	$[F/M_r]^{\#}$
2. Mean	$\bar{Y}_F, \bar{Y}_M$	$[D_o]^{\#}$	$[D_r, D]^{\#}$
3. Standard Deviation	SD_F, SD_M	$[SDR_o]^{\#}$	SDR_r
4. Proportion Selected	$[P_F, P_M]^{\#}$		
5. Correlation with Composite	$[r_F, r_M]^{\#}$		

#Statistics of special interest.

Table D
Effects of Non-normality on Predictions of
Statistics in a Restricted Sample[#]

NELS Test	Predicted (Actual) Values for: Females $\bar{Y}$	Females SD	Males $\bar{Y}$	Males SD	D_r
NELS Scores					
Reading	46.8(46.9)	7.1(2.6)	46.3(46.3)	7.1(2.8)	.07(.19)
Math	68.2(68.2)	9.1(4.8)	70.0(70.1)	9.6(4.5)	-.20(-.40)
History	42.0(42.0)	3.2(1.8)	42.8(42.8)	3.9(1.5)	-.23(-.50)
Science	31.7(31.7)	3.5(1.9)	33.1(33.1)	4.4(1.7)	-.35(-.75)
Composite	5.71(5.71)	.49(.16)	5.83(5.83)	.56(.15)	-.22(-.74)
Normalized Scores					
Reading	49.7(49.7)	5.6(6.2)	48.7(48.7)	5.8(6.8)	.17(.15)
Math	69.6(69.6)	8.3(8.4)	71.8(71.8)	8.5(8.5)	-.27(-.26)
History	42.5(42.5)	3.0(3.5)	44.2(44.1)	3.2(3.6)	-.54(-.47)
Science	31.9(31.9)	3.4(3.6)	34.5(34.4)	3.7(4.2)	-.72(-.65)
Composite	5.80(5.80)	.37(.32)	6.00(6.00)	.38(.36)	-.52(-.57)

[#]All entries refer to the restricted sample which consisted of the top 10% on an equally weighted composite of the four NELS tests.

Table E
Reproducibility of the Effects of Sample Restriction on
Standard Mean Differences (D_r) in the NELS Sample

Measure	Original D_o	Predicted (Actual) Values of D_r in Samples Restricted on the Following Bases:# NELS/HSA $P=.10$		HS Average $P=.50$		ACT/SAT Takers ($P=.60$)	
NELS Reading	.24	.02	(.04)	.15	(.14)	.20	(.20)
NELS Math	-.09	-.42	(-.40)	-.26	(-.25)	-.19	(-.19)
NELS History	-.16	-.60	(-.53)	-.32	(-.31)	-.25	(-.24)
NELS Science	-.31	-.79	(-.75)	-.48	(-.46)	-.39	(-.38)
NELS Composite	-.09	-.61	(-.63)	-.27	(-.25)	-.18	(-.17)
H.S. Average	.31	.46	(.50)	.17	(.17)	.29	(.29)

#The NELS/HSA restricted sample consisted of students who scored in the top 10% on an equally weighted composite of HSA and NELS test composite. The HSA sample consisted of the top 50% on that measure. The test-taker sample consisted of that 60% of the NELS sample who reported having taken either ACT or SAT. HSA and test score distributions were normalized.

Table F

Standard Mean Difference [D_r (D')] for Variable Y in Samples Restricted Through Implicit Selection Using Restricted and Unrestricted Standard Deviations -- Female/Male Ratio = .80[#]

		10% Selected Original Mean Difference					50% Selected Original Mean Difference				
r_{xy}	*SDR*	-.40	-.20	.00	.20	.40	-.40	-.20	.00	.20	.40
.300	1.00	-.38(-.37)	-.18(-.17)	.03(.03)	.24(.23)	.45(.43)	-.36(-.35)	-.15(-.15)	.05(.05)	.26(.25)	.47(.45)
	.95	-.41(-.40)	-.20(-.20)	.00(.00)	.21(.20)	.42(.40)	-.37(-.36)	-.16(-.16)	.04(.04)	.25(.24)	.45(.44)
	.90	-.44(-.42)	-.23(-.22)	-.02(-.02)	.18(.18)	.39(.38)	-.38(-.37)	-.18(-.17)	.03(.03)	.23(.23)	.44(.43)
	.85	-.47(-.45)	-.26(-.25)	-.06(-.05)	.15(.15)	.36(.35)	-.40(-.39)	-.19(-.19)	.01(.01)	.22(.21)	.43(.41)
.500	1.00	-.39(-.35)	-.17(-.15)	.06(.05)	.28(.25)	.51(.45)	-.34(-.31)	-.12(-.11)	.10(.09)	.31(.29)	.53(.49)
	.95	-.44(-.39)	-.22(-.19)	.01(.01)	.23(.21)	.46(.41)	-.36(-.33)	-.14(-.13)	.07(.07)	.29(.27)	.51(.47)
	.90	-.49(-.44)	-.27(-.24)	-.04(-.04)	.18(.16)	.40(.36)	-.39(-.35)	-.17(-.15)	.05(.05)	.27(.25)	.49(.45)
	.85	-.55(-.49)	-.33(-.29)	-.10(-.09)	.12(.11)	.35(.31)	-.41(-.38)	-.19(-.18)	.03(.02)	.24(.22)	.46(.42)
.700	1.00	-.42(-.33)	-.16(-.13)	.10(.07)	.36(.27)	.62(.47)	-.33(-.28)	-.09(-.08)	.15(.12)	.39(.32)	.63(.52)
	.95	-.51(-.39)	-.25(-.19)	.01(.01)	.27(.21)	.53(.41)	-.37(-.30)	-.13(-.10)	.12(.10)	.36(.30)	.60(.50)
	.90	-.59(-.46)	-.33(-.26)	-.07(-.06)	.19(.14)	.45(.34)	-.40(-.33)	-.16(-.13)	.08(.07)	.32(.27)	.56(.47)
	.85	-.68(-.53)	-.42(-.33)	-.16(-.13)	.10(.07)	.36(.27)	-.44(-.37)	-.20(-.17)	.04(.03)	.28(.23)	.52(.43)
.900	1.00	-.53(-.31)	-.18(-.11)	.17(.09)	.52(.29)	.87(.49)	-.34(-.24)	-.06(-.04)	.23(.16)	.52(.36)	.80(.56)
	.95	-.68(-.39)	-.33(-.19)	.02(.01)	.37(.21)	.72(.41)	-.40(-.28)	-.11(-.08)	.18(.12)	.46(.32)	.75(.52)
	.90	-.82(-.47)	-.47(-.27)	-.12(-.07)	.22(.13)	.57(.33)	-.45(-.32)	-.17(-.12)	.12(.08)	.41(.28)	.69(.48)
	.85	-.98(-.56)	-.63(-.36)	-.28(-.16)	.07(.04)	.42(.24)	-.51(-.36)	-.22(-.16)	.06(.04)	.35(.24)	.63(.44)
.999	1.00	-.71(-.29)	-.23(-.09)	.26(.11)	.74(.31)	1.22(.51)	-.37(-.22)	-.04(-.02)	.29(.18)	.62(.38)	.96(.58)
	.95	-.93(-.38)	-.45(-.18)	.04(.02)	.52(.22)	1.00(.42)	-.43(-.26)	-.10(-.06)	.23(.14)	.56(.34)	.89(.54)
	.90	-1.16(-.48)	-.67(-.28)	-.19(-.08)	.29(.12)	.77(.32)	-.50(-.31)	-.18(-.11)	.15(.09)	.48(.29)	.81(.49)
	.85	-1.40(-.58)	-.91(-.38)	-.43(-.18)	.05(.02)	.53(.22)	-.58(-.35)	-.25(-.15)	.08(.05)	.41(.25)	.74(.45)

[#]Tabled entries are D_r and (D'). *SDR* is the female/male standard deviation ratio in variable Y; r_{xy} is the correlation with selection composite X.

Table G

Standard Mean Difference [D_r (D')] for Variable Y in Samples Restricted Through Implicit Selection Using Restricted and Unrestricted Standard Deviations -- Female/Male Ratio = 1.00#

		10% Selected Original Mean Difference					50% Selected Original Mean Difference				
r_{xy}	*SDR*	-.40	-.20	.00	.20	.40	-.40	-.20	.00	.20	.40
.300	1.00	-.42(-.40)	-.21(-.20)	.00(.00)	.21(.20)	.42(.40)	-.41(-.40)	-.21(-.20)	.00(.00)	.21(.20)	.41(.40)
	.95	-.44(-.43)	-.24(-.23)	-.03(-.03)	.18(.17)	.39(.37)	-.42(-.41)	-.22(-.21)	-.01(-.01)	.19(.19)	.40(.39)
	.90	-.47(-.46)	-.27(-.26)	-.06(-.06)	.15(.14)	.36(.34)	-.44(-.43)	-.23(-.23)	-.03(-.03)	.18(.17)	.39(.37)
	.85	-.50(-.49)	-.30(-.29)	-.09(-.09)	.12(.11)	.33(.31)	-.45(-.44)	-.25(-.24)	-.04(-.04)	.17(.16)	.37(.36)
.500	1.00	-.45(-.40)	-.22(-.20)	.00(.00)	.22(.20)	.45(.40)	-.44(-.40)	-.22(-.20)	.00(.00)	.22(.20)	.44(.40)
	.95	-.50(-.44)	-.28(-.24)	-.05(-.04)	.17(.16)	.40(.36)	-.46(-.42)	-.24(-.22)	-.02(-.02)	.20(.18)	.41(.38)
	.90	-.55(-.49)	-.33(-.29)	-.10(-.09)	.12(.11)	.35(.31)	-.48(-.44)	-.26(-.24)	-.05(-.04)	.17(.16)	.39(.36)
	.85	-.61(-.54)	-.38(-.34)	-.16(-.14)	.07(.06)	.29(.26)	-.51(-.46)	-.29(-.26)	-.07(-.06)	.15(.14)	.37(.34)
.700	1.00	-.52(-.40)	-.26(-.20)	.00(.00)	.26(.20)	.52(.40)	-.48(-.40)	-.24(-.20)	.00(.00)	.24(.20)	.48(.40)
	.95	-.60(-.46)	-.34(-.26)	-.08(-.06)	.18(.14)	.44(.34)	-.52(-.43)	-.28(-.23)	-.03(-.03)	.21(.17)	.45(.37)
	.90	-.69(-.53)	-.43(-.33)	-.17(-.13)	.09(.07)	.35(.27)	-.55(-.46)	-.31(-.26)	-.07(-.06)	.17(.14)	.41(.34)
	.85	-.78(-.60)	-.52(-.40)	-.26(-.20)	.00(.00)	.26(.20)	-.59(-.49)	-.35(-.29)	-.11(-.09)	.13(.11)	.37(.31)
.900	1.00	-.70(-.40)	-.35(-.20)	.00(.00)	.35(.20)	.70(.40)	-.57(-.40)	-.29(-.20)	.00(.00)	.29(.20)	.57(.40)
	.95	-.84(-.48)	-.49(-.28)	-.14(-.08)	.21(.12)	.56(.32)	-.63(-.44)	-.34(-.24)	-.05(-.04)	.23(.16)	.52(.36)
	.90	-.99(-.57)	-.64(-.37)	-.29(-.17)	.06(.03)	.41(.23)	-.68(-.48)	-.40(-.28)	-.11(-.08)	.18(.12)	.47(.32)
	.85	-1.15(-.66)	-.80(-.46)	-.45(-.26)	-.10(-.06)	.25(.14)	-.74(-.52)	-.45(-.32)	-.17(-.12)	.12(.08)	.41(.28)
.950	1.00	-.97(-.40)	-.48(-.20)	.00(.00)	.48(.20)	.97(.40)	-.66(-.40)	-.33(-.20)	.00(.00)	.33(.20)	.66(.40)
	.95	-1.19(-.49)	-.70(-.29)	-.22(-.09)	.27(.11)	.75(.31)	-.73(-.44)	-.40(-.24)	-.07(-.04)	.26(.16)	.59(.36)
	.90	-1.41(-.58)	-.93(-.38)	-.45(-.18)	.04(.02)	.52(.22)	-.80(-.48)	-.47(-.28)	-.14(-.08)	.19(.12)	.52(.32)
	.85	-1.65(-.68)	-1.17(-.48)	-.69(-.28)	-.20(-.08)	.28(.12)	-.88(-.53)	-.54(-.33)	-.21(-.13)	.12(.07)	.45(.27)

#Tabled entries are D_r and (D'). *SDR* is the female/male standard deviation ratio in variable Y; r_{xy} is the correlation with selection composite X.

Table H

Standard Mean Difference [D_r (D')] for Variable Y in Samples Restricted Through Implicit Selection Using Restricted and Unrestricted Standard Deviations -- Female/Male Ratio = 1.25#

		10% Selected Original Mean Difference					50% Selected Original Mean Difference				
r_{xy}	*SDR*	-.40	-.20	.00	.20	.40	-.40	-.20	.00	.20	.40
.300	1.00	-.45(-.43)	-.24(-.23)	-.03(-.03)	.18(.17)	.38(.37)	-.47(-.45)	-.26(-.25)	-.05(-.05)	.15(.15)	.36(.35)
	.95	-.48(-.46)	-.27(-.26)	-.06(-.06)	.15(.14)	.35(.34)	-.48(-.47)	-.27(-.27)	-.07(-.07)	.14(.13)	.34(.33)
	.90	-.51(-.49)	-.30(-.29)	-.09(-.09)	.12(.11)	.33(.31)	-.49(-.48)	-.29(-.28)	-.08(-.08)	.13(.12)	.33(.32)
	.85	-.54(-.52)	-.33(-.32)	-.12(-.12)	.09(.63)	.29(.28)	-.51(-.49)	-.30(-.29)	-.09(-.09)	.11(.11)	.32(.31)
.500	1.00	-.51(-.45)	-.28(-.25)	-.06(-.05)	.17(.15)	.39(.35)	-.53(-.49)	-.31(-.29)	-.10(-.09)	.12(.11)	.34(.31)
	.95	-.56(-.50)	-.33(-.30)	-.11(-.10)	.11(.10)	.34(.30)	-.56(-.51)	-.34(-.31)	-.12(-.11)	.10(.09)	.32(.29)
	.90	-.61(-.55)	-.39(-.35)	-.16(-.15)	.06(.05)	.29(.25)	-.58(-.53)	-.36(-.33)	-.14(-.13)	.08(.07)	.29(.27)
	.85	-.67(-.59)	-.44(-.39)	-.22(-.19)	.01(.01)	.23(.21)	-.60(-.55)	-.39(-.35)	-.17(-.15)	.05(.05)	.27(.25)
.700	1.00	-.62(-.47)	-.36(-.27)	-.10(-.07)	.16(.13)	.42(.33)	-.63(-.52)	-.39(-.32)	-.15(-.12)	.09(.08)	.33(.28)
	.95	-.70(-.54)	-.44(-.34)	-.18(-.14)	.08(.06)	.34(.26)	-.67(-.55)	-.43(-.35)	-.18(-.15)	.06(.05)	.30(.25)
	.90	-.78(-.60)	-.52(-.40)	-.26(-.20)	-.00(-.00)	.26(.20)	-.70(-.58)	-.46(-.38)	-.22(-.18)	.02(.02)	.26(.22)
	.85	-.87(-.67)	-.61(-.47)	-.35(-.27)	-.09(-.07)	.17(.13)	-.74(-.61)	-.50(-.41)	-.26(-.21)	-.02(-.01)	.22(.19)
.900	1.00	-.87(-.49)	-.52(-.29)	-.17(-.09)	.18(.11)	.53(.31)	-.80(-.56)	-.52(-.36)	-.23(-.16)	.06(.04)	.34(.24)
	.95	-1.01(-.58)	-.66(-.38)	-.31(-.18)	.04(.02)	.39(.22)	-.86(-.60)	-.57(-.40)	-.28(-.20)	.00(.00)	.29(.20)
	.90	-1.16(-.66)	-.81(-.46)	-.46(-.26)	-.11(-.06)	.24(.14)	-.91(-.64)	-.63(-.44)	-.34(-.24)	-.05(-.04)	.24(.16)
	.85	-1.31(-.75)	-.96(-.55)	-.61(-.35)	-.26(-.15)	.09(.05)	-.97(-.68)	-.69(-.48)	-.40(-.28)	-.11(-.08)	.18(.12)
.999	1.00	-1.22(-.51)	-.74(-.31)	-.26(-.11)	.23(.09)	.71(.29)	-.96(-.58)	-.62(-.38)	-.29(-.18)	.04(.02)	.37(.22)
	.95	-1.44(-.60)	-.96(-.40)	-.47(-.20)	.01(.00)	.50(.20)	-1.02(-.62)	-.69(-.42)	-.36(-.22)	-.03(-.02)	.30(.18)
	.90	-1.67(-.69)	-1.19(-.49)	-.70(-.29)	-.22(-.09)	.27(.11)	-1.10(-.66)	-.77(-.46)	-.43(-.26)	-.10(-.06)	.23(.14)
	85	-1.91(-.79)	-1.43(-.59)	-.94(-.39)	-.46(-.19)	.03(.01)	-1.18(-.71)	-.84(-.51)	-.51(-.31)	-.18(-.11)	.16(.09)

#Tabled entries are D_r and (D'). *SDR* is the female/male standard deviation ratio in variable Y; r_{xy} is the correlation with selection composite X.

Table I
College Admissions Tests and Matching 12th Grade
Test Categories Used in Predicting Effects of
Sample Restrictions on Particular Tests

12th Grade Category#	ACT Tests	College Board Tests
Language Use (6)	English	Test of Standard Written English
Reading (10)	Reading	
Vocab./Reasoning (8)		PSAT-Verbal SAT-Verbal
Math Concepts (13)	Mathematics	PSAT-Math SAT-Math
Natural Science (7)	Natural Science Science Reasoning	
Social Science (3)	Social Studies	

#Number of tests in each category is shown in parentheses.