The Relevance of Group Membership for Personnel Selection: A Demonstration Using Bayes' Theorem
Miller, Edward M The Journal of Social, Political, and Economic Studies; Fall 1994; 19, 3; ProQuest pg. 323

The Relevance of Group Membership for Personnel Selection: A Demonstration Using Bayes' Theorem

Edward M. Miller
University of New Orleans

Groups such as races, sexes, ethnic groups, and age classes are shown to be in general relevant to problems of selection for school and employment. A Bayesian approach to problems of selection is developed. The mean and standard deviation of the distribution of abilities in the candidate's group (race, sex, age etc.) constitute prior information. Additional information is provided by the candidate's test scores. These two can be combined with the aid of Bayes' theorem to obtain a posterior estimate. It is shown that group membership will normally be relevant to selection, with a higher test score being required of those belonging to the lower scoring groups. This implies a fundamental conflict between non-discrimination (not using group membership for selection) and merit selection. The framework developed is then used to show the circumstances in which use of group membership is relevant to selection. Evidence is presented that groups do differ in various attributes relevant to vocational success, including intelligence, literacy, personality, and criminality.

Bayes' Theorem

This journal has repeatedly discussed the technical and ethical issues raised by the existence of groups (races, sexes, ethnic groups) that frequently differ in abilities and other job-related characteristics (Eysenck 1991, Jensen, 1992; Levin, 1990, 1991). This paper is meant to add to that discussion by providing mathematical proof that consideration of such groups is, in general, necessary in selecting the best employees or students.

It is almost an article of faith that race, sex, religion, national origin, or similar classifications (which will be referred to here as groups) are irrelevant for hiring, given a goal of selecting the best candidates. The standard wisdom is that those selecting for school admission or employment should devise an unbiased (in the statistical sense) procedure which predicts individual performance, evaluate individuals with this, and then select the highest ranked individuals. However, analysis shows that even with statistically unbiased evaluation procedures, group membership may still be relevant. If the goal is to pick the best individuals for jobs or training, membership in the group with the lower average performance (the disadvantaged group) should properly be held against the individual. In general, not considering group membership and selecting the best candidates are mutually exclusive.

Three definitions will be used:

> (1) "Non-discrimination" is selection which does not take into account a particular characteristic of the individual being considered (such as race, sex, age, national origin, etc.).
> (2) "Merit Selection" is an endeavor to select the best qualified individual. In the terminology introduced by Hunter and Schmidt (1976), merit selection corresponds to unqualified individualism and non-discrimination to qualified individualism.
> (3) "Ability" here refers to the characteristics sought by the selecting employers or schools, or to the characteristics and interests used in advising. It includes not only ability narrowly defined, but also characteristics such as motivation, honesty, etc.

One of the implications of this paper is that common statements taking the form of "Hiring shall be based on ability irrespective of race (or sex, national origin, religion, handicapped status, marital status, sexual preference, etc.)" are at best ambiguous, and at worst illogical. Logically proper statements are, "The best qualified candidates shall be selected without preference for any group (but taking into account group membership to the extent it is relevant)." or "No consideration of group membership shall be permitted (even when it is necessary to select the best candidate)." In practice, antidiscrimination rules appear to have been sold to the public on the basis of the first statement, but administered on the basis of the second statement. Indeed, not only has consideration of group membership been forbidden even when relevant, but there appears to be a tendency to forbid consideration of any characteristics that might be a surrogate for group membership, or even correlated with it (such as test scores). Rational discussion would be greatly facilitated if participants would state which policy they are advocating.

Proof by Bayes' Theorem

The relevance of group membership is clearly shown by an application of Bayes' Theorem. If f(t) is the probability density function for the ability distribution among the candidates, and the distribution of estimated ability (here referred to as e) given the true ability (symbolized by t) is f(e/t), the distribution of ability given the estimated ability f(t/e) is the product of these two probability density functions divided by the density function for the estimated abilities. The proof is by direct substitution into Bayes' theorem (for Bayes' theorem, see Dyckman et al. [1968, pp. 484-489] or any standard statistics text):

$$f(t/e) = f(t)\, f(e/t) \,/\, f(e) \qquad \textit{Equation 1}$$

Note that f(t) enters into this equation. In general, the probability distribution of abilities among the candidates selected depends on the probability distribution among the candidates being considered. In particular, the mean of the distribution or the final estimate (what is referred to in statistics as the posterior estimate) of the candidate's ability depends on which group the candidate belongs to, and the distribution of abilities within that group. Group membership matters if there are differences in the ability distribution among the different groups. The same argument applies to other characteristics such as personality.

While the above argument clearly leads to a controversial conclusion, it is merely an application of an argument that had been developed for the selection of capital projects (Miller, 1978, 1985), and then extended to personnel selection (Miller, 1980) but without mention of groups. In these contexts it occasioned little controversy. Surprisingly, the above simple point has been missed outside of the technical psychometric literature (which will be discussed later), although several models have been developed in which rational behavior results in different standards for different groups (Aigner and Cain, 1977; Arrow, 1972; Borjas and Goldberg, 1978; Darity, 1989; Phelps, 1972; Schwarb, 1986, Smith, 1978). Also, Epstein (1992, pp. 40 and 240) briefly mentions Bayes' Theorem.

The Special Case of the Normal Distribution

A particularly interesting case occurs if the errors in evaluation of candidates from a particular group are normally distributed, and the distribution of abilities among this group is normal. Human abilities appear normally distributed. It is plausible (from the law of large numbers) that the errors are also normally distributed. Under these conditions, the distribution of true abilities given the estimated abilities (test results) will also be normally distributed with the parameters of the normal distribution easily calculated (see Dyckman, Smidt, & McAdams [1968, p. 486] or other Bayesian texts.)

Let Mp = the mean ability of the group of candidates (the prior mean),
Me = the mean of the distribution of the ability of the candidate given the data about him (excluding information about the group he is a member of)
Mt = the expected value for the ability of the candidate given the estimate (the posterior mean)
sp = the standard deviation of true ability for the population of candidates (the prior estimate)
se = the standard deviation of the estimated ability
st = the standard deviation of the ability of the candidate given the estimate

With this notation,

$$Mt = ((Mp/sp2) + Me/se2)/(1/sp2+1/se2) \qquad \textit{Equation 1}$$

$$Mt = (Mpse2 + Mesp2) / (se2+ sp2) \qquad \textit{Equation 2}$$

And

$$1/st2 = 1/sp2+1/se2 \qquad \textit{Equation 3}$$

The above shows the mean and the standard deviation of the distribution of the ability of the candidate given his estimated ability. The expected ability is of course the mean of the posterior distribution. Equation 2 gives the posterior distribution mean. It is a weighted average of the population mean for a particular group (the prior distribution) and the candidate's estimated ability. The candidate's estimated ability will be referred to as the test score. The evaluation procedure may not involve a written test. Instead, it may be an interview, review of previous performance, or a reference check. The reciprocal of the square of the expected standard deviation of the test error will be referred to as the precision.

In plain English, once a test score has been obtained, the best ability estimate will depend on the average ability of the candidate's group. Thus, if the goal is to select the best candidates, it will be necessary to consider group membership, and the mean ability of the candidate's group. The general effect is to move the ability estimate for each candidate towards his group's mean ability. When trying to select the best candidates (who will usually have evaluations above the mean for their group), the estimate for each candidate should be lowered by an amount that depends on mean and standard deviation for his group, and the estimate's precision.

While this is a conclusion that will bother many, it is one derived by straightforward mathematics. In general, only under special conditions will seeking the best candidates be consistent with disregarding group membership.

If the candidate comes from a "low scoring" group (remembering that what is relevant is the characteristics of the candidates being considered), he should have a higher estimated ability (test score, if the estimates are quantitative) than that required of a member of a "high scoring" group. The above presumes the cut-off score (the minimum score of those hired) is above the group mean. The adjustment towards the group mean lowers the candidate's score. In the cases where the cut-off score is below the group mean (such as where the goal is merely to screen out a small percentage who will prove inadequate), adjustment towards the group mean will raise the estimated score of the individual.

In advising, the logic is the same, although the ethical objections may not be as strong. Presumably candidates' best interests are served if they are given advice based on the best possible estimates of their abilities and interests. If group membership helps in doing this, many might accept using it even though they otherwise oppose its use for hiring or admission purposes.

Groups Differing Only in Score

As a warm-up on a non-controversial topic to see how Bayes' theorem provides new insights, consider the case where groups differ only in ability. Imagine the goal is to hire the workers who will make the fewest errors. Errors are random, but workers differ in their error rates. Work samples (revealing the number of errors made in a one hour test) have been obtained. It is desired to use these samples to estimate how the candidates will do if employed. The work sample appears to be an unbiased measure of the candidates' long-term performance. Naturally, it is only a sample, but the distribution of the sample, given the true performance, is known. What prediction can be made about his performance? For discussion imagine the applicant's score is far above average. Suppose also, the distribution of job performance in the applicant population is also known. What job performance can be predicted?

One might just take the known relationship between test (work sample) performance and job performance, and then argue that the predicted job performance would be that obtained by solving (for job performance) the equation relating test performance to long-term job performance. For instance, suppose the number of errors in a one hour work sample is known to be an unbiased estimate of the workers' error rates during their employment. One would be tempted to estimate the average long-run error rate for a worker with a score of x on the sample as being x.

However, the above procedure would be wrong. It ignores the information about the distribution of the applicants' abilities. Some applicants benefit from good luck, and their scores overstate their true long-run performance. Others suffer from poor luck, and their score understates their true long-run performance. At first glance, it might appear that the two effects would cancel each other out, and it could be presumed that any given candidate was as likely to have benefited from luck as to have suffered from it.

However, as is implied by Bayes' theorem, such canceling frequently won't happen. Even if the expected value of the errors (luck) is zero for the set of all candidates, the expected value of the errors conditional on a candidate having been selected is not zero. The set of those obtaining any given score will include some for whom the score accurately reflects long-run performance, some who benefited from luck (and hence who did better on the test than their long-run performance would justify), and some who did worse. Among those with above average test scores, there will be more who benefited from luck than who suffered from it. For our candidate with the above average score, the best estimate of his long-run performance is obtained by adjusting his score downwards. The predicted mean can be calculated from Bayes' theorem. In more general terms, the application of Bayes' theorem calls for reducing the estimated performance of the high scoring, and raising that of the low scoring. The scores are regressed to the mean.

To understand intuitively the direction of the effect being discussed, consider Figure 1. The distribution of applicants' true abilities is shown. However, these are measured only with an error.

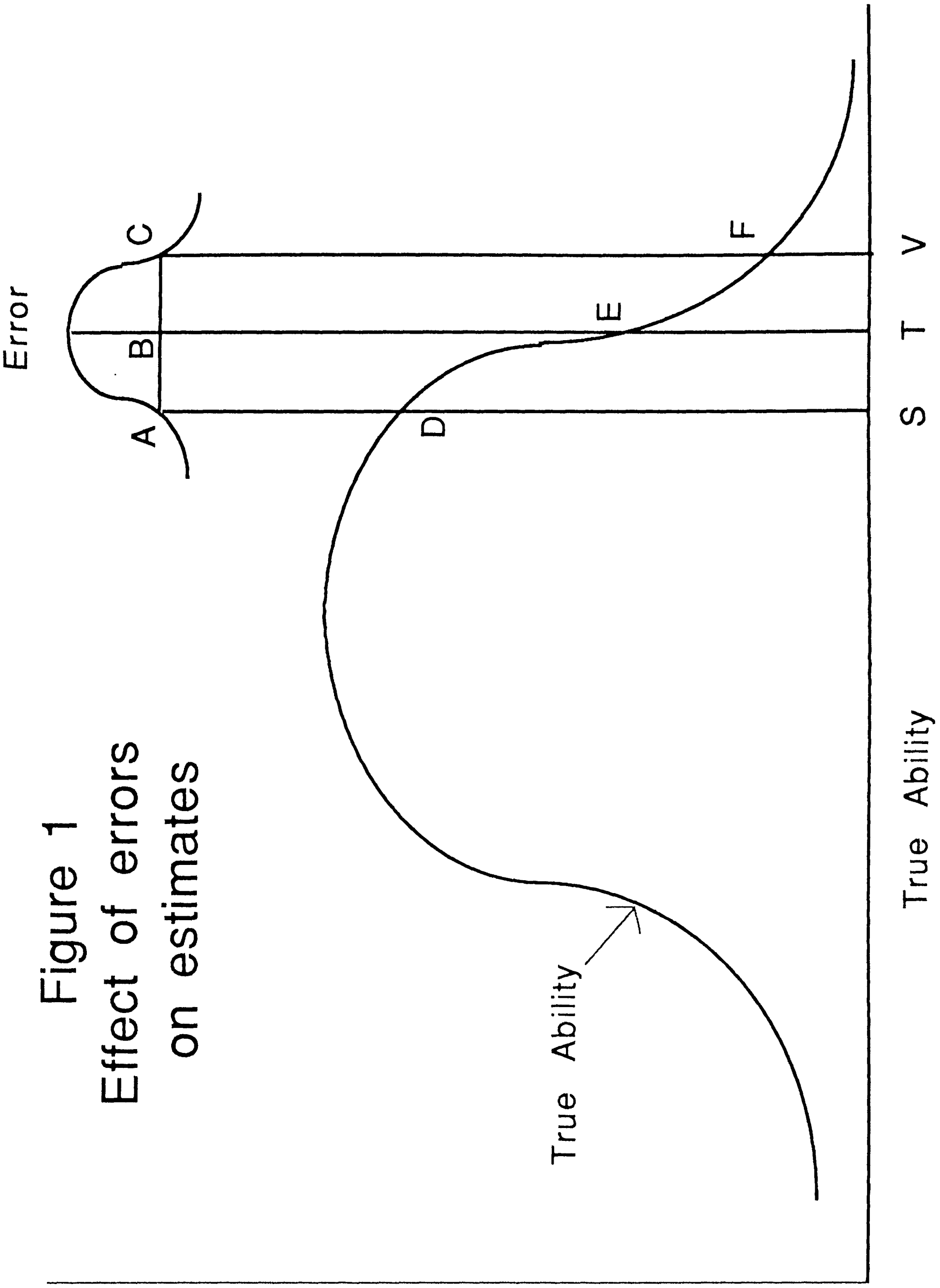

Figure 1
Effect of errors
on estimates

The error corresponds to the small curve shown (presumed to be a symmetrical distribution). Imagine the score reported for a candidate is T. This true value is above the distribution's mean, suggesting the individual is of unusually high ability. However, his ability is measured with error. Consider the error of magnitude AB. An equal error on the high side is BC. With a symmetrical error distribution, the two errors are equal, one high and one low. With a score of T, one realizes the true score could be S (T - AB) or V (T + AB). From the shape of the true value distribution, the probability of the value S must be greater than the probability of the value V. Thus, if an error of absolute value equal to AB or BC has been made, the probability that the true value is S must exceed the probability that the true value is V. The effect of errors of this magnitude is to cause an overestimation of true value.

The argument can be repeated for all possible values for the magnitude of the error. In each case, the conclusion is that the probability of the error causing an overestimate is positive. The conclusion is that the errors lead to an overestimate of the true value. If desired, the magnitude of this overestimate could be calculated, although the above heuristic argument should show why it can be presumed to be positive when one is selecting candidates whose abilities are estimated to be above the average for their group. (The argument is symmetrical for candidates testing below the group mean).

A Work-Sampling Example

Consider the widespread academic goal of selecting professors who will produce many papers. The major source of information about a candidate is the average number of papers per year produced in the previous dozen years. This can be taken as providing information about the number to be produced in the future, but one that contains considerable sampling error (but the distribution of the errors will be presumed known). There is also information about the average number of papers produced by the group the person belongs to, say males or females. The best estimate of the candidate's future productivity is an optimally weighted average of his or her historical productivity and the average productivity of each group.

The weights for the two productivities depend on their relative precision. The general effect will be to adjust the observed rate of paper production for the candidates towards the mean for their groups. The relative weights depend on how much information about future productivity there is in the candidate's historical productivity, and how much is in the group data.

Unless the rate of historical productivity provided perfect predictions of the candidate's future productivity, the group averages would be relevant. Of course, if by some accident, the two groups' means were the same, knowing group membership would add no information. Even in this case, the best estimates for the posterior mean requires adjusting the observed means towards the group mean.

This example can be applied to the case of men and women scientists. Cole (1979, p. 63) reports that after twelve years the average male scientist has produced eight papers, while the average female scientist has produced only three. More recently, Broder (1993) showed that, even after controlling for other relevant variables, female economists have published fewer papers in top journals. Similar results have been found for psychology (Cohen & Gutek, 1991) and academic psychiatry (Reiser, Sledge, Fenton, & Leaf, 1993). During a short period, the observed output of any single scientist is a very imprecise measure of the long-run output of that scientist. Thus, it is necessary to adjust the observed output towards the average for the scientist's group. This adjustment will normally raise the estimated future output of male scientists relative to that of female ones. Thus, where the observed output is only a poor estimate of future output, group membership can significantly improve the precision of estimates.

Related Psychometric Discussions

How does the conclusion reached above about the relevance of groups membership relate to discussions in the technical psychometric literature?

At least some psychometricians have been aware of the relevance of group membership. Hunter and Schmidt (1976) point out that differences in group means will typically lead to differences in intercepts. Jensen (1980, p. 94) points out that the best estimate of true scores is obtained by regressing observed scores towards the mean, and that if there are two groups with different means, the downwards correction for the high scoring individuals will be greater for those from the low scoring group. Kelley (1947, p. 409) put it as follows: "This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates, one based upon the individual's observed score, X1, and the other based upon the mean of the group to which he belongs, M1. If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa", although he may not have been thinking of demographic groups. Cronbach, Gleser, Nanda, and Rajaratnam (1972) discuss the problem of deducing universe scores (essentially true scores in traditional terminology) from test data, recognizing that group means will be relevant. They even display an awareness that, since blacks normally score lower than whites, the logic of their reasoning calls for the use of higher cut-off scores for blacks than for whites (see p. 385). Mislevy (1993) also displays an awareness that group means are relevant, although he feels it would be unfair to use them.

In general, the relevance of group membership has been known to the specialist psychometric community, although few outside the community are aware of the effect. Thus, the contribution of Bayes' theorem is to provide another demonstration, one that those outside the psychometric community may be more comfortable with.

When are Group Means Relevant?

For identical standards to be appropriate, the two groups' means and standard deviations must be identical, and the distribution of errors in the "test" must have the same mean and standard deviation for both groups (i.e., abilities must be equally well estimated for both groups).

There are several reasons why the ability distribution of groups may differ, including:

1. There is a real difference within the total population between the two groups. Since this is a very controversial proposition, it will be discussed later.
2. Within the total population, there are no differences between the groups, but for some reason there are differences among those choosing to apply. Several mechanisms can produce this.
   a. The abilities of the members of the groups who apply for certain jobs may differ because of differences in the other opportunities open to them. For instance, if there is affirmative action in favor of blacks by some employers, the pool of high ability blacks will be depleted, and an employer who does not care about race per se will find that there is a