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,

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

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

And

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

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.

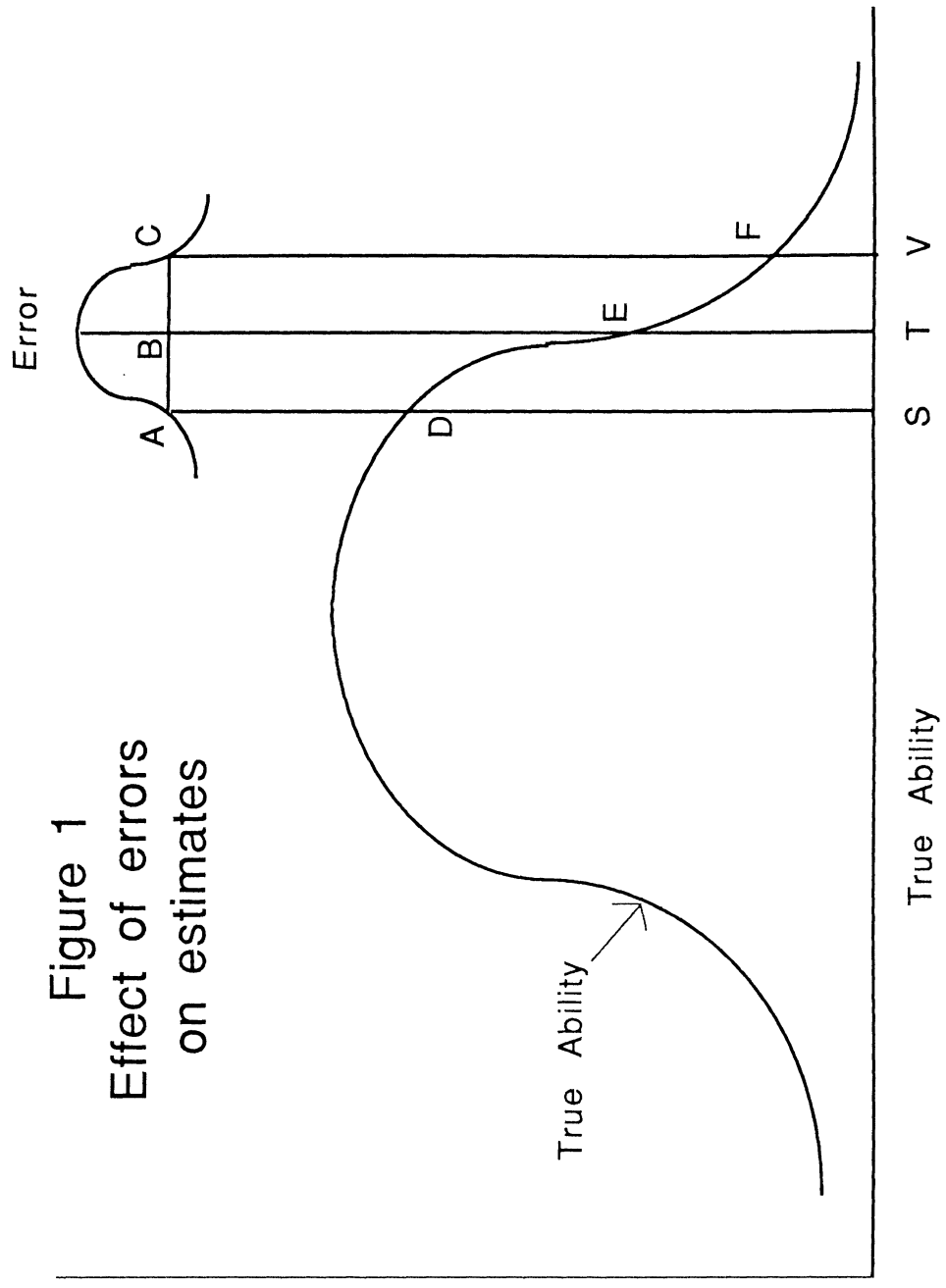

Figure 1
Effect of errors
on estimates

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

lower percentage of high quality blacks among those blacks applying to him, than there is of high quality whites among those whites applying.

b. Differences in tastes may exist. Members of some groups may simply prefer different attributes in their jobs. This is very likely for males versus females. A job involving child care may appeal more to women than to men. A low paying child care job may attract only men who can't get other jobs, but may attract some females who simply prefer taking care of children to other, better paying jobs open to them. Women are known to give more weight in their job search to considerations such as opportunity to serve others, and pleasant co-workers, while men give more weight to status and pay. For certain jobs (especially some with a strong public service component), the only male applicants may be those who can't get jobs with higher status or pay, while some female applicants simply wanted the opportunity to serve, or the work environment. In these cases there will be a sex difference in the applicants' ability distributions, even if population-wide sex differences are lacking.

c. For some reason, the members of one group who apply for a certain job may be less talented than the members of the other group, even though group means do not differ. Suppose there was truly no discrimination, but there was a widespread belief that there was discrimination such that only an exceptionally qualified black could get certain jobs. In this case, the black applicants would be in general better qualified than the white ones.

Notice that in any of these cases, the group means would be relevant information. This of course implies that group membership is relevant. Notice that in the last case, the groups may have equal ability in the whole population, but if a belief in discrimination exists, group means are likely to be relevant. Thus, laws against discrimination (in the sense of considering group means) are likely to be consistent with merit selection only in the case where there is believed to be no discrimination. Of course, if there is believed to be no discrimination, one may ask what the purpose of anti-discrimination laws is?

Role of Precision in Estimates

Equation 2 shows that the weight given to the group mean in predicting a candidate's performance depends on the reciprocal of the standard deviation squared (a quality referred to as the precision) of the initial estimates about the true value. Often, one is trying to select candidates whose performance is superior to that of their group. Typically, only a minority of candidates will be selected. In such a case, candidates from groups for which precise estimates can be made should have an advantage. In plain English, preference should be shown to candidates whose future performance can most easily be predicted.

This result should not prove surprising. To take the extreme example, the best that can be done in the absence of suitable test data is to assume that candidates have their group's average ability (or the average for the whole population if nothing is known about their group). In general, the best scoring members of a group will be better than the average applicant from other groups. Thus, one would expect to do better by taking the best of the candidates about whom there is knowledge than by picking randomly from a group about whom there is no knowledge (other than a group mean).

It is very likely that much of what is called "'prejudice' results from nothing more than the intuitive application of Bayes' Theorem. An employer tries a few members of a particular group. He finds their average performance to be much worse than members of other groups. He recognizes that a few members of the poor performing group perform well. But with no method to discover who these are, the optimal solution is simply to hire no one from the poor performing group. Until after World War I, there were no standardized intelligence tests to discover cheaply the most intelligent members of a poor performing group. Even today, good standardized tests are lacking for many important characteristics.

Imagine two groups of candidates, one consisting of five candidates with whom one has had personal experience, and a hundred believed to be similar in ability, for whom there are only resumes. Picking the highest ranked from those personally known is likely to succeed better than picking the best of the hundred on much lower quality information. Certainly, hiring from within and selecting from professional acquaintances are common.

One often has better information about members of one ethnic or religious group than another. Very often the group the hiring

officer is from is better known to him. The hiring officer may believe the mean abilities of the various groups are equal. However, if he is looking to hire only a few from the applicants in each group, he will adjust downwards more the estimates from the group whose candidates he has less information about. In this case, group membership is relevant because of differences in the precision of the "test" data, rather than because of any differences in the group means. It should be noted that a belief that one group is inferior is not logically necessary for an employer to use higher criteria for members of a particular group. In popular terms, discrimination need not be justified by prejudice.

Experience in similar jobs will usually improve the quality of information about an individual. Such previous experience will usually be greatest for older candidates, increasing the precision of the evaluation. This should give some advantage to the well qualified aged. Of course, those not well qualified will just have their deficiencies made more apparent. Here is a case where precision may depend on what is a prohibited item of information under U.S. law (being over 40). Here age enters into optimal decision-making not because it is correlated with ability, but because it is a surrogate for the precision of estimates.

Tests of ability for different groups often differ in precision. It is widely believed that many written mental tests are culturally biased. Certainly many informal procedures are. Often the most useful questions vary with group membership, and, for reasons of administrative feasibility, only one or a few sets of questions can be used. The most useful set will be that which works best for the majority group. It is plausible that these tests will be less precise for other groups. Notice that in the typical case where only a minority of the candidates are being selected, this implies that the adjustment towards the mean should be less for members of the majority group. The above argument implies that members of minority groups will (and should) be at a disadvantage in competing for most jobs.
However, if selection is by minority group members skilled at evaluating other group members, it may happen that superior members of that group find themselves at an advantage in competing for such positions. Thus, members of ethnic communities should have an advantage within that community. This advantage holds only if they are above average. If the candidates are below the group mean, greater precision in estimating their true ability merely highlights

their inferiority, and makes it harder for them to find employment within the group.

When Group Membership is Unimportant, Repetitive Tasks

So far, this paper has merely shown that group membership will usually be relevant. There remains the question of how important it actually is. This is a big question which can only be touched on here (and is certainly an important topic for research). Theory can give some guidance as to the circumstances in which group membership is relevant.

Theory shows (see Equation 2) that the mean of the posterior distribution will be a weighted average of the means of the prior distribution (i.e., the group mean) and the mean for the test, with the weights being the relative precisions. If individual performance can be accurately predicted from test data, group membership will be of only minor importance (although it will be of no importance only if there is no uncertainty about future performance).

In particular, if the attribute of importance can be accurately measured at low cost, group membership will add little information and should be given little weight. For instance, women may be superior to men in clerical speed and accuracy, and hence make better typists. There could also be racial differences in typing ability. However, typing speed and accuracy are easy to measure. Thus, the optimal weight for sex or race in selection is small (although probably not zero).

In general, if a job consists of a small number of easily learned tasks repeated throughout a day, it will be possible to sample these tasks. Such measurements will have high precision. The optimal weight for group membership will be small. Performance in many (perhaps even most) occupations depends on such easily sampled tasks. Unfortunately, these occupations are often the lower paying ones.

In many such cases it will not be thought worth the trouble (and political heat) of computing the correct weighted average of the test score and the group mean. The test score will just be used. In these cases it should be remembered that if a tie-breaker is needed, the best guess is normally that the candidate from the best scoring group is the best choice (except when hiring from below average candidates). This is opposed to the frequently used procedure of breaking

ties by taking the candidate from the lower scoring group (which is typically the underrepresented group).

Group Membership in Conjunction with Non-Repetitive Tasks

Where full performance on a job is achieved only after extensive training (or on the job experience), performance cannot be established by sampling. The ability of an applicant for admission to a school to learn is hard to sample directly, although measuring (by sampling) what has already been learned is possible. Tests do exist which measure the ability to learn (intelligence tests). These are normally imperfect guides to job or school performance, and hence the weight that should be appropriately given to group membership decreases as the tests become better predictors of true job or school performance.

Estimating Low Probability Events

It is hard to measure precisely the probability of making a very bad mistake or causing an accident. Observation of performance during a short test, or performance in a previous job, may not be of much use. This is because the relevant events occur too infrequently for the probability to be estimated from historical data. Group membership may be the best information available. This is likely in evaluating probabilities such as that of giving secrets to a foreign power, committing another serious crime, causing a serious accident (such as an oil spill), or defaulting on a loan. A controversial example of making decisions affecting national security on the basis of national origin, which has been discussed in this journal, was the internment of many Japanese in World War II (Murphey, 1993).

Where candidates have the ability to impose very high costs by single actions (e.g., raping a fellow student), estimating the probability of such infrequent actions from past experience will be impossible. The only source of useful information may be group membership. Certain forms of handicap may have this property. One way to avoid having a tanker captain drink too much alcohol and run aground in Alaska, causing a major spill, may be to avoid hiring those with a history of alcoholism, even "recovered" alcoholics. However, in the US, alcoholism is considered a disease (which it may be), and it is illegal to discriminate against those with a disease. Unfortunately, the percentage of ex-alcoholics who again take up drinking is much greater than the percentage of those who have never been alcoholics.

There is currently no way of telling which ex-alcoholics will again take up drinking and which won't. Thus starting with a strong "prior" about alcoholics and ex-alcoholics appears correct.

A particularly important case occurs in selecting faculty members or scientists to do highly original research. It is not known how to predict the probability of a Nobel prize-winning insight. Clearly this event is too infrequent to predict from its previous occurrence in the same individual. (If this were the requirement, no junior faculty would ever be hired.) The best rule is to utilize prior information about the groups from which such discoveries have previously come, such as being from a prestigious graduate school, possession of a graduate degree, previous discoveries, etc. A characteristic known to be associated with a low probability of making major discoveries (Lehman, 1953), being past middle age, is one that cannot legally be considered in hiring in the United States. It is possible that other characteristics on the forbidden list may be useful predictors (especially sex, or attributes related to culture, such as religion or national origin). The example given earlier of the large differences in production of scientific papers between men and women suggests sex will frequently prove a useful predictor for scientific productivity. Weyl (1989) has shown that the level of accomplishment in many fields varies with national origin, making it likely that national origin provides relevant information for estimating the probability of accomplishment.

A frequent problem in selection is to screen out dishonest or criminally inclined applicants. It is plausible that honesty will be more likely in candidates that believe that dishonesty is punished after death (i.e., certain religious beliefs) than in candidates whose religious views present no obstacle to dishonesty. Refusing to consider religion can make avoiding dishonesty much harder. Similar effects could occur with other forms of group membership relevant to dishonesty. Consider strength of community ties, or ability to leave the country to avoid prosecution (i.e., nationality or national origin). It is also known that blacks have very high arrest rates and imprisonment rates (Wilson and Herrnstein, 1986; Jaynes and Williams, 1989, pp. 458-461; Levin, 1990, 1991; Rushton, 1994), a fact which might lead a reasonable man to believe that dishonesty was more common in that group. According to mothers' reports (3,049 children) of black ten-and eleven year-old boys, 30.7% were not truthful or told lies (Tuddenham, Brooks, & Milkovich, 1974, as reported by Vernon,

1982). Corresponding figures were 23.9% of the Chicano, 10.8% of the whites, and 8.8% of the Orientals. It is likely that ethnic differences were understated since mother's probably compared their children with other children they knew, who would tend come from the same ethnic group. It is also known that males are much more likely to commit crimes than females are (Rutter & Rutter, 1993, pp. 178-185; Wilson & Herrnstein, 1986). Given the difficulty of getting information about honesty and criminal inclinations, this group information should carry some weight.

Personality Measurements are Typically Low Precision

Much of the problem in predicting low probability events is not measuring ability, but measuring personality. In many selection situations, the key question is not ability to do the job (i.e., what performance will be done when the candidate is motivated) but questions of personality and character. Will the candidate try? Will he consistently show up for school or work? Will she antagonize customers or fellow students? Will he be honest? Does the student's personality match the job she is training for? Personality tests may lack precision in a selection application (partially because candidates can deduce how to answer questions, and answer appropriately), although studies show they can be useful (Schmitt, Gooding, Noe, & Kirsch, 1984; Baehr & Orban, 1989; Barrick & Mount,1991; Tett, Jackson, & Rothstein, 1991; Furnham 1992). With low precision tests, group membership is likely to be useful. In such cases, a higher weight for group membership will be appropriate. In many cases, no attempt is made to measure personality variables at all, even though they are relevant to job and school performance. Group membership then becomes the best guide to personality variables. Of course, if personality can not be measured very precisely, with or without group membership, it should have only limited weight in hiring decisions.

There are differences between various ethnic groups in personality (although it is debated the extent to which they are genetic). Vernon (1982) documents the greater introversion in Oriental peoples. Some difference in behavior are observed at birth (white babies were more excitable while Chinese were more immutable) showing there is some genetic component (Freedman, 1974). The Tuddenham study of mothers' reports mentioned above stated that 41.3% of Chicano boys flare up and get mad easily, versus 28.7% of the blacks, 25.6% of the whites, and only 13.2% of the Orientals.

Assuming differences in such traits carry on into adulthood, as they appear to, one would give high weight to group membership, since reliable individual information about the tendency to flare up is unlikely to be available (candidates are unlikely to admit to this on applications and the behavior is less likely to be exhibited in a brief interview).

The sexes differ in personality (Gilligan, 1982; Moir & Jessel, 1992; Wilson 1989). Sociobiology predicts that because of the different roles in reproduction played by the different sexes there will be real sex differences in behavior (Symons, 1979; Buss, 1994; Ridley, 1993). A prediction of sociobiology is that males will be more oriented toward achieving status, because extra status helps them gain access to more mates, while females benefit relatively little from status because they cannot appreciably increase their number of offspring just by obtaining high status. This greater male desire for status has been argued to explain why men usually achieve the highest status positions in any profession or society (Goldberg, 1973). Given the difficulties in accurate personality assessment, the use of sex is very likely to improve the accuracy of predictions for jobs for which personality is important.

Perhaps the case where the optimal inclusion of a forbidden variable would make the biggest difference would be sex. Females have a greater interest in child care. This could justify a "prior" that a female applicant would do a better job in child care, or in teaching the young. Most women have very strong maternal instincts, and give top priority to caring for their children. This results in many women dropping out of the labor force for a few years after the birth of children. Attempting to forecast who will drop out of the labor force for child care purposes, or who will reduce their work effort, without using sex will likely be less successful than using forecasts that include sex as a direct variable. It is likely that much of the difference in number of publications between male and female academics reflects female decisions to give priority to child care over publishing. Much of the remainder probably results from the males' greater drive for status. Since this decision is a direct result of sex, it is logical that better estimates of future publications will be obtained by making use of a sex variable than without one. Likewise, if the goal of an admissions program for graduate school is to train those most likely to contribute a number of years of work in a professional field, it can be predicted that more of the women will drop out, at least for a few

years, for child care, than will the men. The inclusion of a sex variable will pick up this effect.

Sex differences may be relevant in advising. There is evidence that males and females differ in their occupational goals. Indeed, Levin (1987) argues that most of the difference in occupational distribution between men and women is explainable by men and women having different goals. If this is true, when using instruments given to both males and females, it would be desirable to adjust the estimates towards the means for each sexual group to get the best possible estimate of the true value for their interests. For this reason alone, two individuals who answered a questionnaire identically might be given somewhat different advice depending on their sex. This is a result of the mathematics and does not involve unjustified stereotyping.

Evidence that Group Differences Exist

The literature on group differences is too large for a comprehensive survey to be attempted here. This paper will present just enough evidence that differences in group means, or in group standard deviations, exist (or can be are argued to exist) often enough to make the Bayesian mathematics presented above more than an intellectual curiosity. Unfortunately, there have been extreme efforts to keep the evidence about group differences from becoming known (Pearson, 1991; Eysenck, 1991).

A standard finding is that average intelligence does differ by ethnic group and race. Orientals appear more intelligent than Caucasians (Lynn, 1987, 1991a; Vernon,1982: Rushton, 1994). To avoid the charge that members of one's own ethnic group are normally evaluated as more intelligent than members of other groups, note that the researchers here were Caucasian. Black performance is normally about one standard deviation below white performance (Jensen, 1980; see also Gottfredson, 1986a, 1986b; Herrnstein, 1990; Herrnstein & Murray, 1994; Levin, 1990, 1991; Lynn, 1991a; Osborne & McGurk, 1982; Seligman, 1992; Shuey, 1966). More precisely, the best estimates place the average U.S. black IQ at 82 (Jensen, 1993), with a standard deviation of 12 (which is below the white standard deviation). Incidentally, the fact that the black standard deviation is less than the white one increases the weight to be given to the group mean when estimating the intelligence of a black.

While it is commonly contended that this is because tests are biased against blacks, the evidence is that the major written admission tests and employment tests (especially those that test intelligence or aptitude) are not biased against blacks, and probably not against other English-speaking groups. This conclusion is based on the work of Jensen (1980) and others (Hartigan & Wigdor, 1989; Reynolds & Brown, 1984; Wigdor & Garner 1982; Hunter & Schmidt, 1982; including the Scheuneman, 1987, and Shepard, 1987, dissents). For instance, evidence is that tests of cognitive ability predict as well for blacks as for whites (Schmidt, 1988). While I find Jensen convincing, a poll of expert opinion showed a belief that intelligence tests were "somewhat biased" against American blacks (Synderman & Rothman, 1988, p. 117).

While it is commonly presumed that the sexes are equal in intelligence, Lynn (1994) has recently assembled evidence that adult males have about a 4 IQ point advantage over adult females. Of course, given the precision with which IQ can be estimated, adjustment for this difference would have relatively little impact.

Although there is some evidence that college grade point averages of women are underpredicted, much, if not all, of the underprediction appears to be due to women disproportionately taking the easier courses (i.e., avoiding math-based courses) (see McCornack & McLod, 1988; Elliot & Strenta, 1988; Young, 1991).

Those who are not convinced by these authors' work (perhaps because of arguments by Kamin [1974], Flynn, [1980]; or Ceci, [1990]) can treat the discussion in this paper as purely a logical exercise on the implications of differences should they exist. Perhaps the exercise will be applicable to another case where the reader does believe group differences exist.

Intelligence is closely related to literacy and performance on the types of tasks taught in school. There appear to be large differences between whites and blacks in their ability to read with comprehension, manipulate documents, and engage in mathematical reasoning. An Educational Testing Service study (Kirsch and Jungeblut, 1986) reported on a standardized literacy test administered to a large sample of black and white adults. A high level of performance (375 points) that 10.8% of the whites reached, was reached by only .7% of the blacks. For the quantitative test, 11.5% of the whites reached the equivalent level, versus .8% of the blacks. Given the importance of

these "literacy" skills in high status occupations, the proper "prior" is that blacks would perform much worse than whites.

However, if legally allowed (see Epstein,1992; Gottfredson & Sharf, 1988; Gottfredson, 1988; Hartigan and Wigdor, 1989; Welch, 1989; Wigdor and Garner, 1982 for a discussion of American legal restrictions prior to the Civil Rights Act of 1991, and that act for the current law), testing for such skills would be relatively cheap, and the level of precision high. Thus, there is likely to be considerable information from testing available. In such cases, the additional information on literacy from considering race would be small, although inclusion of race should still improve predictions.

Similar results appear on achievement tests in schools, and on entrance to college and graduate schools (Humphreys, 1988). At the University of California's Medical School (over admission to which the famous Bakke case was fought) there was a very large gap in the Medical College Admission Test scores of those admitted in the regular program, and those admitted in a special program designed to give minorities a share roughly proportional to their percentage in the population (thus implying that selection was from similar percentiles of the population). "The average percentile in which the regularly admitted scored on the verbal section of the MCAT was 81, the specially admitted, 46; the average on the quantitative section for the regularly admitted was 76, the specially admitted, 24; the average on the science section for the regularly admitted was 83; the specially admitted, 35; the average was on the general information section for the regularly admitted was 69, the specially admitted, 33" (Eastland & Bennett, 1979, p. 8). Given the absence of information available to the general public on the competence of individual physicians, rational behavior would give considerable weight to race in choosing a physician. Incidentally, such use of race by an individual purchasing a service appears to be legal, as well as prudent.

Specific employment tests routinely show lower performance by blacks. Epstein (1992, 218-222) describes how in the first test developed by the New York City Police Department for entry-level officers blacks got 7.8% of the passing grades although they were 16.7% of the applicants. This was thrown out on disparate impact grounds (i.e., it failed blacks at a higher rate, and was presumed discriminatory). After a revision of the test was prepared at a high cost, and again thrown out by the courts, a third version was prepared which had a passing rate of 1.6% for blacks (4.4% for Hispanics), but

11% for whites. In spite of such extreme differences in passing rates, (which prevented New York from accepting the outcome of this test), the above logic implies that the appropriate passing scores for blacks should have been higher than for whites, to correct for the possibility that blacks were merely benefitting from chance.

Wynter (1994) reported that in a study of supervisors' ratings of 3600 white managers and 500 black ones, the black managers tended to get lower grades. If blacks are indeed performing more poorly, the above logic calls for reducing supervisors' ratings downwards to allow for their inaccuracies. Notice here that arguing that ratings are inaccurate, which black exponents would likely do, would actually increase the strength of the case for giving more weight to group membership. Of course, a showing of anti-black bias would call for adding points to the blacks, but such a showing has not been made.

There are known to be racial and ethnic differences in disease resistance (Brues, 1990). The mechanism is known in a few cases (the sickle cell trait protects against malaria). Group membership may be the only guide to disease resistance. One could easily imagine conditions, especially in military operations, where it might be desirable to assign individuals on the basis of group affiliation so as to minimize their chances of being killed or injured. For instance, blacks are more vulnerable to frostbite (Brues, 1990, p. 180), but are better able to resist malaria, yellow fever, and other tropical diseases.

The Case for Merit Selection

It may be useful to make the case for merit selection here. Those affected by a personnel decision are the applicants and their employers. Suppose applicants are from two different racial groups. While the members of one race would be made happier if their representative were selected, the opposite would be true if the other were selected. For a non-racist there is no reason to presume that one race should be preferred over the other. Here, a non-racist is considered someone who believes that members of one racial group are not to be preferred over another just because of race. In general, there is no way to know how getting the job will affect the happiness of particular individuals (but see below), or for preferring one race over another on this basis. Thus, in general, a non-racist cannot base decisions on the candidate's welfare.

However, an employer does have a basis for choosing. He will be better off if the candidate selected does the job better. Frequently, of

course, the employer will be a public agency. In this case the benefit of having the job better done will extend to the public. Even if one does not care about private profits, better employees serve the public better. Also, lower prices are likely since fewer man-hours are required when the employees are of higher quality. Thus, with the welfare of the candidates providing no basis for choice, choice should be based on which candidate is expected to do the better job. The same reasoning applies to admission decisions, since giving the training to the best qualified applicants eventually should result in better job performance. For instance, society probably benefits from having the best qualified individuals trained to be doctors, since that should result in more patients correctly diagnosed and treated.

Actually, it may be an overstatement to assume that the utility an individual finds in a job does not depend on ability. A certain percentage of those who get jobs come to regret it. Such regret is especially likely for those who, after getting the job, are fired for incompetence. Among those who do the job well enough to retain it, the pleasure they experience from the job may depend on how well they do it. Those who do a job well generally enjoy it more than those who do it poorly, and experience less of the anxiety that results from poor performance.

Likewise, students who flunk out may regret ever having been accepted. Sowell (1991, p. 110), for instance, has pointed out that affirmative action programs can set black students up for failure at high-prestige universities, even though they could have been successful students at the typical state university. Such failure could affect them for life. At the author's school, when admission was available for all high school graduates, virtually all students who needed remedial work in both English and mathematics failed to graduate. Now minimal standards eliminate most such individuals. Admitting such individuals through affirmative action probably does them no favor. Thus, even from the perspective of applicants, minimizing failure through merit selection is often desirable. As noted above, this requires consideration of group membership.

In general, the question of whom to select depends on the utility function of the selector, and if this includes a forbidden variable such as race, sex, age, or surrogates for the same such as minority status, diversity, previous disadvantage, etc., a strong case can be made for using a selection model with a utility function that makes these

variables explicit (Gross & Su, 1975; Peterson, & Novick, 1976). Group membership enters into most such models.

Admittedly, the idea of using group membership may offend some people's idea of justice, but if that is the case they should make that argument explicitly, and not merely assert that group membership is irrelevant. That considering group membership is immoral is a proposition that can be, and has been, debated. Levin (1992, 1994) has made some interesting arguments that it is proper for the police to consider group membership, even when used to investigate differentially black individuals for criminal activity, although of course that proposition has been disputed (Adler, 1993; Corlett, 1993; Cox, 1993; Pojman, 1993; Thomas, 1992).

In considering these questions of whether using group membership improves selection, the key question is whether there are differences in group means and variances, not what causes these differences. The argument is equally strong whether the differences are of environmental origin or genetic origin. However, it is likely that many of the differences between sexes, age classes, and races do have a genetic origin, and plausible evolutionary accounts are made for why these differences exist. The evolutionary origins of racial differences in intelligences has been discussed elsewhere by the author (Miller, 1991) and others (Lynn, 1991b; Rushton, 1994). The author has set out in detail an evolutionary theory of racial differences which depends on offspring survival in northern climates requiring male provisioning, and on populations in those climates having evolved traits conducive to such provisioning (Miller, 1994). Many personality and behavioral differences between the races are explained by the theory. In warm climates, where Negroids emerged, male traits conducive to male success in competition for mates emerged. The ability and willingness to fight were among these traits. In modern industrial societies a willingness to fight, and to take risks, produces higher crime rates. Also, in cold climates it was necessary to plan ahead to survive winter, and the ability to defer gratification was selected for. Many criminals act as if they had difficulty in deferring gratification. An alternative theory to explain race differences has been proposed by Rushton (1994). However, the evolutionary mechanism proposed appears implausible (Miller 1993), although Rushton & Ankney (1993) disagree.

Implications for Test Theory

Predicting a Criterion from Test Scores

Suppose measures of performance are known for a large sample of individuals whose test scores are also known. The distribution of test scores given performance can be determined, perhaps by regressing the test score on the criterion. (Notice the contrast with the usual procedure.)

By solving this equation for performance as a function of test scores, a prediction for performance can be obtained. Should this prediction be used? No. In general it will overpredict the performance of those with high scores, and underpredict the performance of those with low scores. The set of those with high (i.e., above average) scores includes more whose scores were raised by luck than whose scores were lowered by luck. Thus, it overpredicts performance.

There is a better procedure. Given a knowledge (or estimate) of the distribution of abilities (performance) in the applicant population, Bayes' theorem can be used to compute the expected performance given the test score.

Using the regression of the test score on the criterion plus Bayes' theorem is not standard procedure. Indeed, Thorndike (1971, p. 64), in referring to the regression of the test score on the criterion, states: "Generally, this second regression is not of practical value, though it may be of theoretical interest in some contexts." As will be argued, this regression is of practical value whenever the candidates' ability distribution differs from that of the validation population.

The standard procedure is the computationally simpler one of regressing the job performance on the test scores. Then when an estimated score is obtained, one reads off the regression line the estimated long-run performance that corresponds to that score.
If the applicant's distribution is the same as the validation population, this procedure is correct. The regression of the criterion on the test score gives (by definition) the best fitting line.

If the set of applicants with a particular score includes more individuals whose long-run performance has probably been overestimated than applicants whose long-run performance has probably been underestimated, the standard procedure adequately corrects for this, but only if the set of applicants has the same distribution as the set of subjects for which the test is validated.

The Applicant Distribution Differs from the Distribution in the Validation Sample

However, there is a big qualification to the above which has been given inadequate emphasis. In most applications, a tacit assumption is that the distribution of ability among the applicants is the same as in the validation sample (the sample which was used to show the test worked).

Usually, only the best of the applicants for a job or for school entrance have been accepted. It is only the accepted ones for which criterion data is available. In these cases the distribution of true abilities in the validation sample will differ from the distribution in the pool of applicants on which decisions must be made. (This is the well known restriction of range problem. Procedures exist for estimating validities for a whole population. These will not be discussed here.)

The critical point to note is that in most cases the sample used to validate the test has different statistical properties from those of the population the test will be used for. In these circumstances, standard test theory fails to provide a good rationale for regressing the criterion on the test score. At best, this appears a reasonable, but ad hoc procedure.

It seems much more reasonable to explicitly use Bayes' theorem. An estimate of the candidate's performance conditional on his test score is needed. Bayes' theorem can provide this if the distribution of abilities among the candidates can be obtained, and if the distribution of the test scores given the candidate's ability can be obtained. The latter can be obtained by regressing the test score on the criterion. Where the test is an actual work sample, statistical theory will likely provide the distribution for the sample conditional on the criterion.

The harder problem is knowing the candidates' ability distribution. Typically, only data for the candidates actually hired exists. If the distribution of ability is known, this data can be used to reconstruct the full distribution. For instance, the means and standard deviation for the applicant pool could be estimated from the average ability of those hired, the minimum ability found acceptable, and the knowledge that the top x% of the applicants were hired. This would permit the calculation of a table for the expected performance given the test score. If candidates are drawn from different groups, this can be done separately for each. In general, this will result in a different table for each group.

Use of Bayes' theorem provides an explicit role for the distribution of abilities among the set of candidates. One can alter the selection rules as the distribution of applicants changes, or as one estimates it to have changed. Recruiting campaigns may draw better or worse applicants, for instance. One at least has a conceptual framework in which to discuss this problem.

However, in the most common applications of test scores, the outcome is not altered by the assumption that the candidates' ability is distributed as in the validation sample. Typically a test is used to rank applicants. The top applicants are then selected. Fortunately, for a single population or group, adjustments for differing distributions of abilities among the candidates merely changes the estimated abilities for particular candidates, without changing the candidates' rankings. Thus, there is no need for the full Bayesian apparatus where the only need is to rank candidates, and all candidates come from a single group.

As can be seen from the examples discussed here, where candidates are drawn from two different populations, improved results are obtained by using Bayes' theorem to adjust candidates' scores.

Cronbach, Gleser, Nanda, and Rajaratnam (1972, p. 384) have pointed out that many test applications involve content or criterion referenced measurement, such as determining whether a student has mastered the material well enough to proceed to the next instructional unit. For these purposes, the problem is to estimate the student's true mastery, and the best estimate is a weighted average of the subpopulation mean and the test score, as shown above with Bayes' theorem.

Group Membership in Traditional Test Theory

The analysis also indicates the typical conditions in which group membership is useful. Traditional test theory at most provides for empirically adding group membership as a parameter in estimates, possibly as a dummy variable. Alternatively, separate regression coefficients can be estimated for each group. Traditional theory would use statistical tests to decide whether to include group membership as a dummy, or to fit a separate regression line for each group. Unless the probability is less than some traditional amount (often .05), the dummy is set equal to zero, or a single regression line

is used. The procedure appears somewhat ad hoc, and an impression is left that the procedure is rather undesirable.

The above analysis suggests that whenever groups differ in average ability, a group effect is to be expected. The value of a group dummy should not be arbitrarily set at zero just because the coefficient is not statistically significant. If the two groups have different means, group membership will in general be relevant, and a term for group membership should appear in the equations.

Using the apparent criterion of whether differing regression coefficients exist, Hunter, et al. (1984, p. 79) concluded, "Given the cumulative research findings available today, there is no longer any separation between qualified and unqualified individualism that is empirically relevant. That is, although the positions can be separated philosophically, the empirical situations that would result in a difference in outcome do not occur." As long as it is assumed that the major reason for differing regression coefficients is that group membership is serving as a surrogate for some uncontrolled characteristic (which it may be doing), whether there is a difference between qualified and unqualified individualism is an empirical issue, refutable by showing that regression coefficients statistically differ no more than they would by chance. The Bayesian analysis shows that a difference in regression lines is to be expected whenever the statistical distribution of abilities differs between groups. The empirical evidence shows that such differences in means between groups occur regularly, creating a presumption that the regression lines will differ.

This makes the philosophical issue a real one, even if today the issue is less important than whether or not certain groups should be given affirmative action preference. If for some reason the lines do not differ when the groups are known to differ (perhaps from other tests), the logical presumption is that the tests are biased at the item level. This can happen through a careful effort to select items a particular group does well on.

A major use of the analysis given here is in showing the fundamental conflict between merit selection and non-discrimination. This would logically lead one who believed in merit normally to oppose anti-discrimination laws. Economists (such as the author), with their traditional professional bias towards rationality, would be expected to choose merit selection.

Of course, one who was in favor of merit selection still might rationally support laws against considering group membership.

Perhaps real life people are so blinded by prejudice that they fail to make optimal decisions, even though such decisions are in their own interest. Such a person (possibly a psychologist with a knowledge of how frequently irrationality, including irrational prejudice, is actually observed) might choose to forego the gains from optimal consideration of group membership for the gains of avoiding sub-optimal use of group membership. Since the arguments about sub-optimal use of group membership information have been frequently made, this paper has concentrated on pointing out that use of group membership will normally be necessary for optimal decisions.

Conclusions

Only under highly unusual conditions should a firm or school seeking the best candidates have the same requirements for members of different groups. These necessary conditions include (1) identical average abilities and distribution of abilities, (2) identical precision of the evaluation procedure (test) in predicting future performance, (3) no tendency for the better qualified members of one group to apply for either the job being offered or for jobs that might compete for the same candidates. The last condition normally requires that the tastes of the different groups be identical with regard to the desirability of different jobs or schools, that selection for schooling and employment not be based on group membership (if it is, there will be depletion effects regardless of whether the use of group membership is rational), and that returns to human capital investment be the same in all groups.

For laws forbidding the consideration of group membership to be consistent with seeking the best qualified applicants, a universal belief that discrimination does not occur is necessary. (Otherwise, self-selection by candidates is likely to make group membership relevant.) Of course, if it is agreed that there is no discrimination, it is hard to imagine why it should be forbidden. It appears that an interest group cannot logically argue both that there is a problem with discrimination, and that outlawing consideration of group membership is consistent with seeking the best candidates.

Of course, rules against consideration of group membership may be justified on other grounds than the irrelevance of group membership. It may be that selection procedures using group membership involve externalities (such as promoting social strife) that justify forbidding their use. It may be believed that a certain group should

be given a greater share of the available jobs or school places, and that forbidding consideration of group membership facilitates this goal. Finally, there is the second best argument that decision makers do not know how to make optimal use of group membership data, and that better decisions result from forbidding use of such data than from permitting it to be incorrectly used.

Of course, if support for rules forbidding consideration of group membership is based on the above considerations, proponents should make these arguments. They should not use the logically incorrect argument that ignoring group membership is necessary for seeking the best qualified candidates. It is not even consistent with that goal. Likewise, if the goal is to help people make the best vocational choices for themselves given imperfect instruments for measuring their interests and aptitudes, group membership is likely to be relevant. Vague comments about "stereotyping" do not alter this mathematical fact.

In general, hiring decisions affect those hired and their employer. If the ethical decision is made that there is no reason for preferring the welfare of one individual over that of another because of his race or other group membership (which seems to be a widely shared belief), it appears that the only basis for hiring will be the benefits to the employer and the general public. This normally requires seeking the best qualified candidates. Since seeking the best qualified candidates is in general inconsistent with equality of opportunity, it follows that the absence of group consideration is inconsistent with equality of opportunity. Hard choices must be made.

References

Adler, J.
1993 Crime rates by race and causal relevance: A reply to Levin. Journal of Social Philosophy, XXIV, 176-184.

Aigner, D. J., & Cain, G. G.
1977 Statistical theories of discrimination in labor markets. Industrial Labor Relations Review, 30, 175-187.

Arrow, H. J.
1972 Some mathematical models of race discrimination in the labor market. In A. H. Pascal, (ed.), Racial Discrimination in Economic Life, Lexington.

Baehr, M. E., & Orban J. A.
1989 The role of intellectual abilities and personality characteristics in

determining success in higher-level positions. Journal of Vocational Behavior, 35, 270-287.

Barrick, M. R., & Mount, M. K.,
1991 The big personality dimensions and job performance: a meta--analysis. Personnel Psychology, 44 (1), 1-26.

Borjas, G. J., & Golberg, M. S.
1978 Biased screening and discrimination in the labor market. American Economic Review, 68(5), 918-922.

Broder, I.
1993 Professional achievement and gender differences among academic economists. Economic Inquiry, XXXI, 116-127.

Brues, Alice M.
1990 People and Races. Prospect Heights: Waveland.

Buss, David M.
1994 The Evolution of Desire. New York: Basic Books.

Ceci, S. J.
1990 On Intelligence ... More or Less, Englewood Cliffs, Prentice Hall.

Cohen, A. G., & Gutek, B. A.
1991 Sex differences in the career experiences of members of two APA divisions. American Psychologist, 46, 1292-1298.

Cole, J. R.
1979 Fair Science: Women in the Science Community, New York: The Free Press. Journal of Social Philosophy, XXIV, 155-162.

Corlett, J. A.
1993 Racism and affirmative action. Journal of Social Philosophy, XXIV, 163-175.

Cox, C. B.
1993 On Michael Levin's 'responses to race differences in crime'. Journal of Social Philosophy, XXIV, 155-162.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N.
1972 The Dependability of Behavioral Measurements, New York: John Wiley & Sons.

Darity, W., Jr.
1989 What's left of the economic theory of discrimination? in S. Shulman, & W. Darity, Jr, (Eds), Question of Discrimination: Racial Inequality in the U. S. Labor Market Middletown, Conn.: Wesleyan University Press. 335-376.

Dyckman, T. R., Smidt, S., & McAdams, A. K.
1969 Management Decision Making Under Uncertainty, London, MacMillan.

Eastland, T., & Bennett, W. J.
1979 Counting by Race. New York: Basic Books.

Epstein, Richard A.
1992 Forbidden Grounds. Cambridge: Harvard University Press.

Elliot, R., & Strenta, A. C.
1988 Effects of improving the reliability of the GPA on prediction

generally and on comparative predictions for gender and race particularly. Journal of Educational Measurement 25, 333-347.

Eysenck, Hans J.
1991 Science, Racism, and Sexism. Journal of Social, Political, and Economic Studies, 16(2), 218-250.

Flynn, James R.
1980 Race, I. Q. and Jensen, New York: Routledge, Chapman, and Hall.

Freedman, D. G.
1974 Human Infancy: An Evolutionary Perspective. New York: Wiley.

Furnham, A.
1992 Personality at Work. London: Routledge.

Gilligan, C.
1982 In a Different Voice, Cambridge: Harvard University Press

Goldberg, S.
1973 The Inevitability of Patriarchy. New York: Morrow.

Gottfredson, L. S.
1986a The g Factor in employment: A Special Issue of the Journal of Vocational Behavior, 29 (3).
1986b Societal consequences of the g factor in employment. Journal of Vocational Behavior, 29, 379-410.
1988 Reconsidering fairness: A matter of social and ethical priorities. Journal of Vocational Behavior, 33, 293-319.

Gottfredson, L. S., & Sharf, J. C.
1988 Fairness in Employment Testing: A Special Issue of the Journal of Vocational Behavior,33 (3).

Gross, A. L., & Su, W.
1975 Defining a "fair" or "unbiased" selection model: a question of utilities. Journal of Applied Psychology 60, 345-351.

Hartigan, J. A., & Wigdor, A. K.
1989 Fairness in Employment Testing. Washington, National Academy Press.

Herrnstein, R. J.
1990 Still an American dilemma. The Public Interest, 98, 3-17.

Humphreys, L. G.
1988 Trends in levels of academic achievement of blacks and other minorities. Intelligence, 12, 231-260.

Hunter, J. E., & Schmidt, F. L.
1976 A critical analysis of the statistical and ethical implications of five definitions of test fairness. Psychological Bulletin, 83(6), 1053-1071.
1982 Ability tests: economic benefits vs. the issue of fairness. Industrial Relations, 21(3), 293-308.

Hunter, J. E., Schmidt, F. L., & Hunter, R.
1979 Differential validity of employment tests by race: a comprehensive review and analysis. Psychological Bulletin, 86(4), 721-735.

Jaynes, G. D., & Williams, R. M., Jr.
1989 A Common Destiny: Blacks and American Society, Washington:

National Academy Press.

Jensen, A. R.

1980 Bias in Mental Testing, New York: The Free Press.

1981 Straight Talk About Mental Tests, New York: The Free Press.

1992 Mental Ability: Critical Thresholds and Social Policy. Journalo f Social, Political, and Economic Studies, 17(2), 171-181.

1993 Black intelligence. In Sternberg, R., Ed. Encyclopedia of Intelligence. New York: Macmillan.

Kamin, L. J.

1974 The Science and Politics of IQ, Hillsdale: Lawrence Erlbaum.

Kelley, T. L.

1947 Fundamentals of Statistics, Cambridge: Harvard University Press.

Kirsch, I. S., & Jungeblut, A.

1986 Literacy: Profiles of America's Young Adults, Princeton: Educational Testing Service.

Lehman, H. C.

1953 Age and Achievement, Princeton: Princeton University Press.

Levin, M.

1987 Feminism and Freedom. New Brunswick: Transaction Books.

1990 Implications of Race and Sex Differences for Compensatory Affirmative Action and the Concept of Discrimination. Journal of Social, Political, and Economic Studies, 15(2), 175-212.

1991 Race Differences: An Overview. Journal of Social, Political, and Economic Studies, 16(2), 195-216.

1992 Responses to race differences in crime. Journal of Social Philosophy, XXIII, 5-29.

1994 Reply to Adler, Cox, and Corlett. Journal of Social Philosophy, 25, 5-19.

Lynn, R.

1978 Ethnic and racial differences in intelligence: international comparisons. in R. Travis Osborne, Clyde E. Noble, & Nathaniel Weyl (Eds.) Human Variation: The Biopsychology of Age, Race, and Sex, (261-283) New York: Academic Press.

1987 The Intelligence of the Mongoloids. Personality and Individual Differences, 8, 813-844.

1991a Race differences in intelligence: a global perspective. Mankind Quarterly, 31, 254-296.

1991b The evolution of racial differences in intelligence, Mankind Quarterly, 32, 99-121.

1994 Sex differences in intelligence and brain size: A paradox resolved. Personality and Individual Differences, 17, 257-271,

McCornack, R., & McLod, M.

1988 Gender bias in the prediction of college course performance. Journal of Educational Measurement 25, 321-331.

Miller, E. M.

1978 Uncertainty Induced Bias in Capital Budgeting, Financial Man-

agement, 7, 12-18.
1980 Personnel Selection in the Presence of Uncertainty, Personnel, September-October, 67-76.
1981 Inconsistency of Non-Discrimination with Merit Selection, Resources in Education, (micro-fiche), September.
1985 Decision-Making under Uncertainty for Capital Budgeting and Hiring, Managerial and Decision Economics, 6, 11-18.
1991 Climate and intelligence. Mankind Quarterly, 32, 127-132.
1993 Could r Selection Account for the African Personality and Life Cycle. Personality and Individual Differences, 15, 665-676.
1994 Paternal Provisioning versus Mate Seeking in Human Populations, Personality and Individual Differences, 17, 227-255.

Mislevy, R. J.
1993 Some formulas for use with Bayesian ability estimates. Educational and Psychological Measurement 53, 315-328.

Moir, A., & Jessel, D.
1992 Brain Sex, New York: Dell Publishing.

Herrnstein, R., & Murray, C.
1994 The Bell Shaped Curve. New York: The Free Press.

Murphey, Dwight D.
1993 The World War II Relocation of Japanese-Americans. Journal of Social, Political, and Economic Studies, 18(1), 93-118.

Osborne, R. T., & McGurk, F. C. J.
1982 The Testing of Negro Intelligence, Volume 2, Athens, Ga.: The Foundation for Human Understanding.

Pearson, R.
1991 Race, Intelligence and Bias in Academe. Washington: Scott Townsend.

Peterson, N. S., & Novick, M. R.
1976 An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3-29.

Phelps, E. S.
1972 The statistical theory of racism and sexism. American Economic Review, 62, 659-661.

Pojman, L.
1993 Responses to crime: A response to Michael Levin and Laurence Thomas. Journal of Social Philosophy, XXIV, 152-154.

Reiser, L. W., Sledge, W. H., Fenton, W., & Leaf, P.
1993 Beginning careers in academic psychiatry for women-"Bermuda Triangle"? American Journal of Psychiatry 150, 1392-

Reynolds, C. R., & Brown, R. T.
1984 Perspectives on Bias in Mental Testing, New York: Plenum.

Ridley, Matt
1993 The Red Queen. New York: MacMillian.

Rushton, J. P.
1994 Race, Evolution and Behavior: A Life History Perspective. New

Brunswick: Transaction Publishers.
Rushton, J. P., & Ankney, C. D.
1993 The evolutionary selection of human races: a response to Miller, Personality and Individual Differences, 15(6), 677-680.
Rutter, M., & Rutter, M.
1993 Developing Minds. New York: Basic Books.
Scheuneman, J. D.
1987 An argument opposing Jensen on test bias: the psychological aspects. in Modgil, S. & Modgil, C. (Eds.) Arthur Jensen: Consensus and Controversy, New York: Falmer, 155-170.
Schmidt, F. L.
1988 The problem of group differences in ability test scores in employment selection. Journal of Vocational Behavior, 33, 272-292.
Schmidt, F. L., & Shepard, L. A.
1987 The case for bias in tests of achievement and scholastic aptitude. in Modgil, S. & Modgil, C. (Eds.) Arthur Jensen: Consensus and Controversy, New York: Falmer, 177-190.
Schmitt, N., Gooding, R., Noe, R., & Kirsch, Michael
1984 Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.
Schwarb, S.
1986 Is statistical discrimination efficient. American Economic Review, 76, 228-234.
Seligman, Daniel
1992 A Question of Intelligence. New York: Birch Lane Press.
Shuey, A.
1966 The Testing of Negro Intelligence, New York: Social Science Press.
Smith, M. M.
1978 Towards a general equilibrium theory of racial wage discrimination. Southern Economic Journal, 45(2), 458-468.
Sowell, T.
1991 Preferential Policies. New York, William Morrow.
Symons, D.
1979 The Evolution of Human Sexuality. New York: Oxford University Press.
Synderman, M., & Rothman, S.
1988 The IQ Controversy, the Media and Public Policy. New York: Transaction Publishers.
Tett, R. P., Jackson, D. N., & Rothstein, M.
1991 Personality Measures as predictors of job performance: a meta-analytic review. Personnel Psychology, 44,, 703-742.
Thomas, L.
1992 Statistical badness. Journal of Social Philosophy, XXIII, 30-41.
Thorndike, R. L.
1971 Concepts of culture-fairness. Journal of Educational Measurement,

8, 63-70.

Tuddenham, R. D., Brooks, J., & Milkovich, L.
1974 Mothers reports of the behavior of ten-year-olds and relationships with sex, ethnicity, and mother's education. Developmental Psychology, 10, 959-995.

Vernon, P. E.
1982 The Abilities and Achievements of Oriental in North America. New York: Academic Press.

Welch, F.
1989 Affirmative action and discrimination. in S. Shulman & W. Darity, Jr. (Eds.) Question of Discrimination: Racial Inequality in the U. S. Labor Market. Middletown, Conn.: Wesleyan University Press. 153-189.

Weyl, N.
1989 The Geography of American Achievement. Washington: Scott-Townsend.

Wigdor, A. K., & Garner, W. R. (Eds)
1982 Ability Testing: Uses, Consequences, and Controversies, Washington: National Academy Press.

Wilson, Glenn
1989 The Great Sex Divide. Washington: Scott Townsend.

Wilson, J. Q., & Herrnstein, R. J.
1986 Crime and Human Nature, New York: Touchstone Books.

Wynter, L. E.
1994 Black managers reject white bosses criticism. Wall Street Journal, February 2), p. B1.

Young, J. W.
1991 Gender bias in predicting college academic performance: a new approach using item response theory. Journal of Educational Measurement 28, 37-47.