FAA-AM-79-3

A COMPARISON OF THREE MODELS FOR DETERMINING TEST FAIRNESS

Mary A. Lewis

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma

January 1979

Document is available to the public through the
National Technical Information Service,
Springfield, Virginia 22161

Prepared for
U.S. DEPARTMENT OF TRANSPORTATION
Federal Aviation Administration
Office of Aviation Medicine
Washington, D.C. 20591

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-AM-79-3		

4. Title and Subtitle	5. Report Date
A COMPARISON OF THREE MODELS FOR DETERMINING TEST FAIRNESS	NOVEMBER 1978
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
MARY A. LEWIS	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, Oklahoma 73125	11. Contract or Grant No.
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	
Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, S.W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplementary Notes

This work was performed under Task AM-C-78/79-PSY-66.

16. Abstract

In addressing the issue of test fairness, the sample size of the minority group is usually very small. Thus, sampling error alone could cause a test to appear unfair to either the majority or minority group, when in fact both groups were sampled from populations with identical test and criterion characteristics. There are three prominent models of test fairness in the dichotomous situation: (a) Thorndike's Constant Ratio model (the ratio of the proportion successful to the proportion selected should be equal for the majority and the minority groups); (b) Darlington's Conditional Probability model (the probability of selection, given that an individual is successful, should be equal for both groups); and (c) Einhorn and Bass' Equal Probability model (the probability of success, given that an individual is selected, should be equal for both groups). The present study explored, using a Monte Carlo technique, the robustness of these models to divergent sample size. This technique allows the generation of normally distributed variables of known means, standard deviations, and intercorrelations. One hundred samples of 1,000 subjects, having predictor/criterion correlations of .3 and .4, were generated. Minority samples of 100 subjects were randomly selected from each large sample and the three models were applied. The models were compared for their robustness to sample size differences, different predictor/criterion correlations, and different selection and success ratios. Results indicated that all three models were equally able to identify test fairness under the conditions specified in the present study. The choice of model to use when evaluating test fairness must remain a subjective one based on the fairness goals of the testing agency and further definition of test fairness by Federal guidelines.

17. Key Words	18. Distribution Statement
Test Fairness Personnel Selection Selection Criteria	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	14	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

A COMPARISON OF THREE MODELS FOR DETERMINING TEST FAIRNESS

I. Introduction.

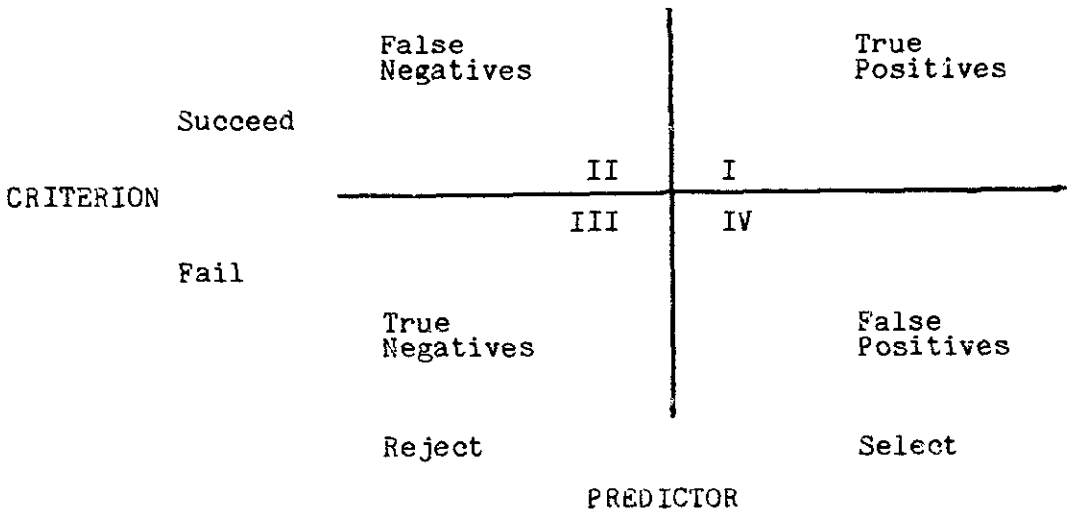
The Uniform Guidelines on Employee Selection Procedures (1978) (9), which were recently adopted by the U.S. Civil Service Commission, the Equal Employment Opportunity Commission, the Department of Justice, and the Department of Labor, state that a selection procedure has an adverse impact if the selection rate for any racial, ethnic, or sex group is less than four-fifths of the rate for the group with the highest selection rate. The guidelines further state that these same rules apply to any employment decision, which can include training, retention, or promotion. The current Air Traffic Control (ATC) training program conducted at the Federal Aviation Administration's (FAA) Academy is a pass/fail program which affects whether or not the trainee will be retained by the FAA in the ATC option. As such, it involves an employment decision and is subject to the standards for validation research and fairness defined by the guidelines.

Although the Uniform Guidelines acknowledge that "the concept of fairness or unfairness of selection procedures is a developing concept," they require that, when feasible, a test must be demonstrated to be fair. The guidelines further specify that "unfairness is demonstrated through a showing that members of a particular group perform better or poorer on the job than their scores on the selection procedure would indicate through comparison with how members of other groups perform." The key concept in this definition of fairness is that performance of a group is compared to the performance of the larger group on both the selection procedures and the job performance measures. If performance is not the same for both groups on both measures, unfairness may exist.

Unfortunately, deciding when "performance is not the same" is not as simple as it may seem. The literature has many articles offering approaches to the evaluation of test fairness. However, these articles seldom deal with the distribution of various fairness indices, nor do they address directly the decision processes involved in deciding whether or not a test is fair. Several authors have found that the major definitions of test fairness lead to conflicting conclusions about test fairness (1,4,7). In addition, Hunter and Schmidt (5) concede that they cannot agree on a definition of test fairness. The available literature offers many methods of evaluating test fairness but little guidance in choosing the most appropriate method.

Most of the models of test fairness define it in psychometric terms. The three major models to be discussed in the present study define fairness in the dichotomous case in which an applicant is either accepted or rejected based on a predictor score and would succeed or fail based on a criterion. Table 1

Table 1. Three Definitions of Test Fairness



CONSTANT RATIO MODEL (CR) - Thorndike (1971) The ratio of the proportion successful to the proportion selected should be equal for both the majority and minority groups.

$$\frac{I_a + II_a}{I_a + IV_a} = \frac{I_b + II_b}{I_b + IV_b}$$

CONDITIONAL PROBABILITY MODEL (CP) - Darlington (1971) The probability of selection, given that an individual is successful, should be equal for both the majority and minority groups.

$$\frac{I_a}{I_a + II_a} = \frac{I_b}{I_b + II_b}$$

EQUAL PROBABILITY MODEL (EP) - Einhorn and Bass (1971) The probability of success, given that an individual is selected, should be equal for both the majority and minority groups.

$$\frac{I_a}{I_a + IV_a} = \frac{I_b}{I_b + IV_b}$$

where a = majority group
b = minority group

graphically depicts this situation and states the three major models of test fairness, verbally and mathematically, in terms of the four cells depicted in the table.

The first model is Thorndike's (8) Constant Ratio model (CR) which states that for a test to be fair, the ratio of the proportion successful to the proportion selected should be equal for the minority and the majority groups. Expressed in terms of the cells in Table 1, the ratio of the sum of the cells I and II to the sum of cells I and IV should be equal for both groups. Darlington's (2) Conditional Probability model (CP) states that a test is fair if the probability of selection, given that an individual is successful, is equal for both groups. In terms of the cells in Table 1, the ratio of cell I to the sum of cells I and II should be equal for both groups. Finally, Einhorn and Bass (3) propose the Equal Probability model (EP) in which a test is considered fair if the probability of success, given that an individual is selected, is equal for both the minority and the majority groups. In terms of the cells in Table 1, the ratio of cell I to the sum of cells I and IV should be equal for both groups. The three models differ in the target groups to which they are "fair." The Constant Ratio model is aimed at insuring that the proportion of applicants selected from both groups is fair. If this model is used, an equitable proportion of applicants from both groups will be hired. The Conditional Probability model is targeted at successful individuals and is intended to insure that an equitable number of successful individuals will be hired. The Equal Probability model is targeted at individuals already hired and is intended to insure that an equitable number of hired individuals will be successful. These models can lead to conflicting conclusions about the fairness of a test. However, there is very little in the literature to describe the distribution characteristics of the three models and how their distributions differ.

The purpose of the present study is to evaluate the distribution of the fairness statistics generated by the Constant Ratio, the Conditional Probability, and the Equal Probability models of test fairness. Since the sample size is, in general, much smaller for the minority sample than for the majority sample, the three fairness indices will be compared for a large sample and a smaller sample across different success ratios on both the criterion and the predictor and also across different correlations of predictor and criterion. Research studies have shown that sampling error leads to an inverse relationship between sample size and correlations (6). It is expected that sampling alone should cause the correlations for the small sample to be higher than corresponding correlations for the large sample. The Constant Ratio model is not sensitive to differences in the correlation of the predictor and criterion, while the Conditional Probability and the Equal Probability models are. It is expected that the Constant Ratio model will be more robust to sampling errors related to sampling size than will either the Equal Probability or the Conditional Probability model.

II. Method.

The data used for analysis in this study were computer generated by using a Monte Carlo technique. This approach allows the generation of a number of normally distributed variables with specified means, standard deviations, and intercorrelations. The technique essentially allows definition of the characteristics of a population and then selects samples from that population. A score of 70 or greater was arbitrarily set as a cut score, scores above 70 were defined as successful for the criterion variable, and scores above 70 were defined as selected for the predictor. Variable means and standard deviations were assigned values such that either 60 percent, 70 percent, or 80 percent of the sample would be above the cut score, and predictor/criterion correlations of .3 or .4 were assigned. Nine variables were generated for this study by using the proportion above 70 and the correlations specified in Table 2. The success rates, selection rates, and predictor/criterion correlations were chosen based on recent experience with the FAA's Air Traffic Control selection and training program. The 18 possible combinations of selection ratio, success ratio, and predictor/criterion correlation described in Table 3 were evaluated.

Table 2. Proportion Above a Score of 70 Assigned Each Variable and Relevant Correlations Input Into Monte Carlo Program

Proportion	Var # [1]	1	2	3	4	5	6	7	8	9
.60	1	X	.3	.4	.3	X	X	X	X	X
.60	2		X	X	.4	X	X	.3	X	X
.60	3			X	X	X	X	.4	X	X
.70	4				X	.3	.4	X	X	X
.70	5					X	X	X	.3	X
.70	6						X	X	.4	X
.80	7							X	.3	.4
.80	8								X	X
.80	9									X

1 The correlations denoted by X were not used in the analysis.

Table 3. All Possible Combinations of Selection Ratio, Success Ratio, and Predictor/Criterion Correlation

	Selection Ratio	Success Ratio	Rxy	x variable	y variable
1	.60	.60	.3	1	2
2	.60	.60	.4	1	3
3	.60	.70	.3	1	4
4	.60	.70	.4	2	4
5	.60	.80	.3	2	7
6	.60	.80	.4	3	7
7	.70	.60	.3	4	1
8	.70	.60	.4	4	2
9	.70	.70	.3	4	5
10	.70	.70	.4	4	6
11	.70	.80	.3	5	8
12	.70	.80	.4	6	8
13	.80	.60	.3	7	2
14	.80	.60	.4	7	3
15	.80	.70	.3	8	5
16	.80	.70	.4	8	6
17	.80	.80	.3	7	8
18	.80	.80	.4	7	9

Each sample that was generated contained 1,000 subjects of which 100 were randomly assigned to the minority group and 900 were assigned to the majority group. Since both the minority and the majority groups were from the same population, the predictors should be equally fair across success ratios, selection ratios, and predictor/criterion correlations. The CP, EP, and CR indices were calculated for the 18 conditions described in Table 2. This process was repeated 100 times.

III. Results.

Table 4 shows the average proportion above a score of 70 and the average intercorrelation matrix obtained across the 100 large samples and the 100 small samples. Table 5 gives the distribution characteristics of three fairness indicators for both the large samples and small samples when the various combinations of selection ratios, success ratios, and predictor/criterion ratios are combined. Table 6 gives the distribution characteristics of the large and small sample fairness indicators when the selection ratio is equal to the success ratio, when the selection ratio is less than the success ratio, and when the selection ratio is greater than the success ratio. Table 7 contains the distribution characteristics of the large and small sample fairness indicators when the predictor/criterion correlation is .3 or .4.

In order to compare the fairness indices for the large and small groups, the indices were expressed first as a ratio of the large group index to the small group index (LG/SM), and then as a ratio of the small group index to the large group index (SM/LG). The distribution characteristics of these indices are described in Table 8.

Table 4. The Average Proportion Above a Score of 70 and the Average Correlation Matrix Across the 100 Large Samples and the 100 Small Samples

For 100 Large Samples[1]

Average Proportion	Var #	1	2	3	4	5	6	7	8	9
.608	1	X	0.31	0.42	0.30	X	X	X	X	X
.603	2		X	X	0.44	X	X	0.31	X	X
.643	3			X	X	X	X	0.43	X	X
.703	4				X	0.34	0.45	X	X	X
.727	5					X	X	X	0.29	X
.712	6						X	X	0.41	X
.308	7							X	0.37	0.42
.806	8								X	X
.816	9									X

For 100 Small Samples[1]

Average Proportion	Var #	1	2	3	4	5	6	7	8	9
.590	1	0.42	0.53	0.32	X	X	X	X	X	X
.583	2		X	0.30	X	X	X	0.42	X	X
.607	3			X	X	X	X	0.47	X	X
.727	4				X	0.23	0.43	X	X	X
.714	5					X	X	X	0.39	X
.700	6						X	X	0.57	X
.780	7							X	0.31	0.44
.730	8								X	X
.802	9									X

1 The correlations denoted by X were not used in the analysis.

Table 5. Distribution Characteristics for the Three Fairness Indicators for the Large and Small Samples

	Mean	SD	Range Lo	Range Hi
CRLG	1.02	.16	.74	1.35
CRSM	1.01	.18	.67	1.49
CPLG	0.77	.07	.63	0.88
CPSM	0.77	.09	.57	0.94
EPLG	0.78	.07	.63	0.88
EPSM	0.77	.09	.57	0.94

Correlation Matrix

	CRLG	CRSM	CPLG	CPSM	EPLG	EPSM
CRLG	1.000	.956	- .821	- .753	.791	.737
CRSM		1.000	- .776	- .787	.758	.755
CPLG			1.000	.886	- .311	- .298
CPSM				1.000	- .324	- .202
EPLG					1.000	.902
EPSM						1.000

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample

Table 6. Distribution Characteristics for the Three Fairness Indicators for Large and Small Samples Comparing Selection Ratio and Success Ratio

Selection Ratio Equals Success Ratio

	Mean	SD	Range Lo	Range Hi
CRLG	1.017	.024	.97	1.08
CRSM	.999	.045	.89	1.11
CPLG	.773	.058	.68	.86
CPSM	.778	.076	.61	.88
EPLG	.786	.055	.69	.86
EPSM	.776	.074	.62	.89

Selection Ratio Is Less Than Success Ratio

	Mean	SD	Range Lo	Range Hi
CRLG	1.194	.081	1.10	1.35
CRSM	1.220	.099	1.00	1.49
CPLG	.703	.046	.63	.77
CPSM	.698	.057	.57	.79
EPLG	.836	.035	.76	.88
EPSM	.847	.045	.73	.94

Selection Ratio Is Greater Than Success Ratio

	Mean	SD	Range Lo	Range Hi
CRLG	.841	.054	.74	.91
CRSM	.825	.068	.67	1.00
CPLG	.836	.035	.76	.88
CPSM	.847	.045	.73	.94
EPLG	.703	.046	.63	.77
EPSM	.698	.057	.57	.79

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample

Table 7. Distribution Characteristics for the Three Fairness Indicators for Large and Small Samples Comparing Predictor/Criterion Correlations

Predictor/Criterion Correlation Equals .3

	Mean	SD	Range Lo	Range Hi
CRLG	1.016	.165	.74	1.35
CRSM	1.013	.182	.67	1.49
CPLG	.761	.074	.63	.87
CPSM	.760	.088	.57	.91
EPLG	.763	.074	.63	.87
EPSM	.758	.087	.57	.91

Predictor/Criterion Correlation Equals .4

	Mean	SD	Range Lo	Range Hi
CRLG	1.019	.145	.78	1.28
CRSM	1.017	.173	.69	1.44
CPLG	.781	.069	.68	.88
CPSM	.789	.082	.62	.94
EPLG	.787	.067	.68	.88
EPSM	.790	.081	.62	.94

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample

Table 8. Distribution Characteristics for the Ratios of the Three Fairness Indicators

	Mean	SD	Range Lo	Range Hi
CR LG/SM	1.01	.05	.88	1.15
CR SM/LG	1.00	.05	.87	1.14
CP LG/SM	1.00	.06	.86	1.20
CP SM/LG	1.00	.05	.83	1.17
EP LG/SM	1.00	.05	.86	1.20
EP SM/LG	1.00	.05	.83	1.17

Correlation Matrix

	CR LG/SM	CR SM/LG	CP LG/SM	CP SM/LG	EP LG/SM	EP SM/LG
CR LG/SM	1.000	- .997	- .554	.544	.448	- .438
CR SM/LG		1.000	.574	- .563	- .426	.416
CP LG/SM			1.000	- .996	.493	- .502
CP SM/LG				1.000	- .502	.513
EP LG/SM					1.000	- .996
EP SM/LG						1.000

where CR is the Constant Ratio model
CP is the Conditional Probability model
EP is the Equal Probability model
LG is the large sample
SM is the small sample

IV. Discussion.

As expected, Table 4 shows that the correlations for the small samples tended to be higher than those for the large samples. It is not surprising that for all three fairness indicators, the small sample groups demonstrated greater variation than did the larger sample groups. The range of the fairness indicator was virtually identical for the CP and EP models, and was a smaller range than that for the CR model. This is to be expected since the CP and EP indices could range only from 0 to 1, while the CR index could range from 0 to infinity.

When the distributions of fairness indicators are examined for the three relationships of selection ratio to success ratio described in Table 6, it can be seen that all three tend to have moderate values when selection ratios are equal; CR and EP have high values when selection ratios are greater than success ratios, while the CP value tends to be higher when the selection ratio is greater than the success ratio. Both CP and EP show the greatest amount of variance when the selection ratio is equal to the success ratio, while CR shows the greatest amount of variance when the selection ratio is less than the success ratio. When the distributions of the fairness indices for the large and small samples are examined separately for correlations of .3 and .4 (see Table 7), all three fairness indicators have lower means and higher standard deviations for the lower correlation.

The fairness indicator ratios described in Table 8 show that the distribution differences observed in Table 5 virtually disappear. The means of these ratios are around 1.00 (as they should be when the test is "fair"); the small standard deviations and the range of the ratios are almost identical for the large group/small group and for the small group/large group indices. It would appear that all three fairness indicators show similar patterns of covariance between the large sample and small sample groups.

Based on the data from the present study, there is no compelling statistical reason to choose any one of the three fairness indicators over the others. The range of the values of the indicators is affected by both the relationship of selection and success ratios, and predictor/criterion correlations. However, while the magnitude of the fairness indicator may vary, the relationship of the fairness indicators for the large and small groups remains about the same, no matter which fairness indicator is used. The three fairness indicators are equally likely to lead the investigator to conclude that a test is fair when the majority and minority groups are chosen from the same population and differences between the groups are due to sampling. Quite frequently, however, this is not the case in the real world. Members of minority and majority groups may be recruited in different ways and may differ dramatically in education, experience, socioeconomic status, and other demographic variables that will affect their performance on the selection devices. The applicants from the majority and minority groups may have different means on the selection tests, and if the means for the minority group are lower than the means for the majority group, then the proportion

selected from the minority applicants could well be less than four-fifths the proportion selected from the majority applicants. If this is the case, then the Uniform Guidelines state that adverse impact has occurred, and the user must demonstrate that the selection test is fair.

The Constant Ratio model could be used at this point to determine if the differential proportion selected for the minority group is compensated for by a differential success rate. If the CR definition of fairness is met, it is unlikely that the selection procedure as defined will be perceived as unfair. The CR model is insensitive to the magnitude of the correlation of the predictor and the criterion, so it would be possible to meet the CR definition of fairness while still selecting majority and minority applicants with vastly different probabilities of success. If this is the case, and if the minority group members selected have a lower probability of success than the majority group members, the minority group members will have a higher attrition rate during the training process than the majority group members. Since the Uniform Guidelines are extended to cover not just selection procedures, but also employment decisions including promotion, referral, retention, and transfer, the user may find that at some point after selection some other employment decision demonstrates adverse impact. If the Equal Probability model of test fairness is used, this problem may be avoided, but unless the regression lines for the minority and majority groups have the same slopes, its use could result in the disproportional selection of one group or the other. The Conditional Probability model could be used to insure that appropriate numbers of successful individuals are selected, but its use too could result in an inequitable selection ratio.

The test user is in a dilemma, as current definitions and practices stand. In order to meet the definition of fairness at the point of selection, the Constant Ratio model may be employed, but use of this model may result in adverse impact and unfairness at some later employment point. The acceptability of the various fairness decision models will no doubt be determined by the courts. In the ideal case, in which the minority and majority samples are selected from the same population and their regression lines are identical, all three models will agree, as they did in the present study. If the test user is in the unpleasant situation in which the models would lead to conflicting conclusions about test fairness, then some corrective action must be taken. If the Equal Probability model indicates test fairness, but the CR and CP do not, then an unfair proportion of successful minorities are being rejected, and a lower cut score may be justifiable. This will occur when the predictor/criterion correlation is higher for the minorities than for the majority. If the Conditional Probability model indicates test fairness, but the EP and CR do not, then the predictor/criterion correlation is lower for the minority than for the majority, and resolution of this problem may require either development of new selection procedures or recruitment of a minority applicant population that more closely resembles the majority sample.

If the use of different cut scores is not feasible, or if the data indicate that the minority applicants differ from the majority applicants in how well their performance can be predicted, the test user could examine recruitment practices to see if efforts could be made to recruit minority applicants who are more like the majority applicants in terms of characteristics related to the probability of success. The most recent version of the Uniform Guidelines emphasizes the role of recruitment and its effect on fairness. This emphasis on recruitment indicates that the effects of recruitment practices on selection and other employment decisions will be a part of the evaluation of the fairness of a selection procedure. Modification of minority recruitment practices could be an effective means of bringing existing selection procedures into compliance with the Uniform Guidelines without necessitating the development of new selection devices.

REFERENCES

1. Breland, H. M., and G. H. Ironson: Defunis Reconsidered: A Comparative Analysis of Alternative Admission Strategies, JOURNAL OF EDUCATIONAL MEASUREMENT, 13:89-99, 1976.

2. Darlington, R. B.: Another Look at "Culture Fairness," JOURNAL OF EDUCATIONAL MEASUREMENT, 8:71-82, 1971.

3. Einhorn, H. J., and A. R. Bass: Methodological Considerations Relevant to Discrimination in Employment Testing, PSYCHOLOGICAL BULLETIN, 75:261-269, 1971.

4. Hunter, J. E., and F. L. Schmidt: Critical Analysis of the Statistical and Ethical Implications of Various Definitions of Test Bias, PSYCHOLOGICAL BULLETIN, 83:1053-1071, 1976.

5. Hunter, J. E., and F. L. Schmidt: Bias in Defining Test Bias: Reply to Darlington, PSYCHOLOGICAL BULLETIN, 85:675-676, 1978.

6. Pace, L. A., and J. L. Mendoza: Increasing Validity With Decreasing Sample Size. Presented at the Annual Meeting of the American Psychological Association, Washington, D.C., September 1976.

7. Sawyer, R. L., N. S. Cole, and J. W. Cole: Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection, JOURNAL OF EDUCATIONAL MEASUREMENT, 13:59-76, 1976.

8. Thorndike, R. L.: Concepts of Culture-Fairness, JOURNAL OF EDUCATIONAL MEASUREMENT, 8:63-70, 1971.

9. Uniform Guidelines on Employee Selection Procedures. Federal Register, Vol. 43, No. 251, p. 38290, August 25, 1978.