The application of a simplified model to assess the fairness of employee selection measures in a sample of white and coloured apprentices

X.C. Birkenbach

Department of Industrial and Organizational Psychology, University of Port Elizabeth, P.O. Box 1600, Port Elizabeth, 6000 Republic of South Africa

A. Allan

Department of Mercantile Law, University of Port Elizabeth, P.O.Box 1600, Port Elizabeth, 6000 Republic of South Africa

Accepted 3 June 1988

While the topic of fairness in personnel selection decisions has received a great deal of attention in the United States of America over the last two decades or so, little work has been done in South Africa in this regard. In view of the possible industrial relations implications of discrimination linked to test unfairness, research needs to be undertaken in this country to examine predictor-criterion relationships for the various race groups who make up the labour force. In the present study, a simplified approach for evaluating test fairness developed by Lawshe (1983) was tested in a sample of white and coloured apprentices. Four selection instruments were related to three criteria of job success. The results of the research seem to verify the results of studies conducted in America, namely, that little evidence has emerged supporting the concepts of differential validity and differential prediction of tests for different race groups. Future research should include samples consisting of black employees in order to compare their test and criterion profiles with those of persons from other racial groups.

Alhoewel daar reeds baie aandag aan die onderwerp van toetsregverdigheid in personeelseleksie in die Verenigde State van Amerika geskenk is, is daar nog min navorsing in hierdie verband in Suid-Afrika gedoen. In die lig van moontlike nywerheidsverhouding-aksies as gevolg van toetsdiskriminasie, moet daar aansienlik meer navorsing gedoen word ten einde toets-kriteriumverhoudings vir die verskillende rassegroepe te ondersoek. In die huidige studie is die model wat deur Lawshe (1983) voorgestel is om toetsregverdigheid te bepaal in 'n steekproef bestaande uit wit en kleurling vakleerlinge geëvalueer. Die verband tussen vier seleksie-instrumente en drie kriteria van werksukses is ondersoek. Die resultate van die studie ondersteun bevindings van navorsing wat in Amerika gedoen is, naamlik, dat daar min indikasie is van die teenwoordigheid van differensiële geldigheid en differensiële voorspelling van toetse vir verskillende groepe. Toekomstige navorsing moet let op die insluiting van steekproewe bestaande uit swart werknemers ten einde toets- en kriteriumprofiele van laasgenoemde te vergelyk met die van ander rassegroepe.

For many decades South Africa has been characterized by two worlds of work, i.e., a skilled one populated primarily by whites and the other semi- and unskilled populated largely by blacks. It is only in recent years that black employees have been entering job categories which were traditionally occupied by whites. In fact, a new buzz word has developed in management circles, namely, 'black advancement' which has triggered a movement in many organizations to actively accelerate the progress of blacks in the work-place. The protected status of white employees in South Africa is rapidly diminishing and the possibility that blacks and whites may compete directly for positions in organizations has now become a reality.

A very urgent question in the minds of many managers today is whether employee selection measures which have been used among white populations will be appropriate for selection of employees from other population groups which make up South Africa's labour force. Owing to the different educational, economic, and socio-political positions of blacks in the South African society, the question posed is whether blacks will achieve comparable scores on psychometric selection measures to their white counterparts (e.g., see Raubenheimer, 1983). If test scores and criterion performance are similar for the different racial groups then the tests can be seen as being fair to both groups. If, on the other hand, test scores for a group (e.g., blacks) are lower than those for another group (e.g., whites) but the *criterion* scores are similar, then the test would be underpredicting black job performance and, consequently, can be seen as discriminating against blacks. Apart from the moral issues involved, this could be construed as an unfair labour practice with concomitant Industrial Relations implications. Furthermore, the misuse of human resources is something the country can ill afford.

The concept of unfair labour practice

The concept 'unfair labour practice' was recently introduced into South African law when a definition to this effect was inserted in the Labour Relations Act 28 of 1956 by section 1(f) of the Industrial Conciliation Act 94 of 1979. The original definition has subsequently been amended by the Industrial Conciliation Act 95 of 1980 and the Labour Relations Amendment Act 51 of 1982.

The definition is very wide and general and, as is the case in respect of other fair employment issues in the Act, the Wiehahn Commission envisaged that the Industrial Court would develop a body of case law which would by judicial precedent contribute to the formulation of guidelines of what would constitute unfair labour practice.

It is anticipated that South African employees will some time in the future raise the question of whether or not the use of certain selection devices constitute an unfair labour practice. In fact, this has already happened. The dispute between SAAWU and Continental China is an example where the company, having dismissed all the employees, offered to reemploy individuals who had been identified as the most efficient and productive workers on the basis of 'objective' tests. The union objected to the tests on various grounds because, *inter alia*, they were based on educational level (Golding, 1985). The 'fairness' issue was not raised in the Industrial Court and at this stage there is no local legal precedent which the Industrial Court could use, should this be alleged. The question arises as to where the court will search for guidance in such a matter. In view of the large number of cases abroad (especially in the United States of America) which deal with this question, it appears logical to look overseas for guidelines in such decisions. This would, of course, be problematical because of differences between the political and socio-economic cultures here and overseas.

In a recent case, Mahlangu vs CIM Deltak, Gallant vs CIM Deltak (1986), the court had to decide whether the use of a polygraph tester (a Mark II Voice Analyzer) as a lie detector was fair under the circumstances. The court held that in the absence of any relevant South African case law on the subject of lie detectors, the decisions of foreign jurisdictions ought to have a strong persuasive influence on the Industrial Court's decision and should serve as guidelines. However, in Bleazard vs Argus Printing & Publishing Co Ltd (1983), the same court warned that 'one should be cautious in relying on foreign sources in interpreting and developing the concept of unfair labour practice'. The reason for this is that the South African legal system and legislation in respect of unfair labour practice may differ from that of the relevant foreign jurisdiction in which the case was decided (Ehlers, 1982).

The Lawshe model of test fairness

Whereas the decision about what constitutes an unfair labour practice involves moral and legal arguments, at least some psychometric evidence can be presented regarding the fairness or unfairness of psychological tests when they are used to make decisions about the classification of people in work organizations. Unfortunately many of the models of test fairness are relatively complex, especially in terms of their statistical underpinnings. An exception is the simple procedure which has been advanced by Lawshe (1983).

According to this model, predictor and criterion scores are converted to standard scores. This is followed by a calculation of the so-called prediction error which is achieved by obtaining the difference between the standardized predictor and criterion scores. The mean prediction error scores for members of each group (e.g., blacks and whites) are then calculated. Finally, the significance of the difference between these means is determined, e.g., by means of a *t* test. If the errors of prediction for the combined groups are evenly spread across the overpredicted and underpredicted quadrants of the regression scattergram, then the mean errors of prediction will be approximately zero. If, however, the mean of one group (e.g., blacks) is significantly smaller than the mean of the other group (e.g. whites), then the test has underpredicted the black group's criterion performance. This can be regarded as evidence that the test has discriminated against this group.

Aim of the study

It is against this backdrop that the present study was launched. While substantial research has been generated in the United States of America over the years in connection with test fairness, there has been a paucity of work here in South Africa. Moreover, as implied earlier, it cannot be assumed that the findings of research carried out in America can summarily be extrapolated to the situation in this country. Given the fact that employees and unions are increasingly challenging decisions which traditionally have been regarded as management prerogatives, it has become very important to assess test fairness in selection and placement decisions. A further problem in this regard concerns the fact that many of the models of test fairness which have been advanced rely on relatively sophisticated statistical analyses which preclude their use by most practitioners. Thus, the aim of the present study was to demonstrate the use of the comparatively simple model of test fairness presented by Lawshe (1983) in a sample of employees in a South African organization.

Method

Subjects

Data were gathered from company records for 52 white and 51 coloured apprentices who were employed in a large manufacturing organization. Because data were available for only 17 black employees, this group was not included in the analysis.

Predictor measures

Four employee selection instruments which were used in the company to select apprentices constituted the independent variables of the study. These are briefly described.

Panel interview: The panel consisted of four to five members who were line managers, personnel office officials and apprentice school tutors. Semi-structured interviews were used to assess applicants' suitability for the positions for which they had applied. The panel allocated a score to each interviewee which was then converted to a percentage . Those persons with the higher scores were deemed to be most suitable for the job.

Blox test: This is a pencil and paper test of spatial cognition which has shown acceptable levels of reliability for black, white, and coloured apprentices. The test contains 45 items.

High Level Figure Classification Test (HLFCT): This is a non-verbal measure of intellectual ability which is

normally used for the selection of workers for jobs requiring abstract conceptual functioning. High reliabilities of this test in black and white samples have been recorded. There are 24 pencil and paper items in the test.

Computation Test: This is a 30 item sub-test of the NIPR Intermediate battery. It consists of arithmetical computation items. The reliability of this test for white respondents seems relatively low but quite acceptable for blacks.

Criterion measures

Two objective criteria and one relatively subjective criterion were used in the study to act as indicators of employee success.

Practical criterion: This measure of job performance consisted of the average of the marks given to apprentices for the practical work they performed, such as making models according to specific dimensions.

Theory criterion: This measure consisted of the average mark which respondents achieved for theory tests which they wrote from time to time.

Supervisor's evaluation: This was a relatively subjective estimate of the apprentices' work performance as made by the supervisors to whom the apprentices reported. Evaluations were made on seven dimensions of work performance (e.g., quality of work, housekeeping, and safety) on a 10-point scale with behavioural anchors. Scores were summed over the eight dimensions to produce a single total score for each apprentice. The internal consistency reliability of this criterion as assessed by means of Cronbach's alpha is 0,91. (Owing to practical problems in the work setting it was not possible to calculate the reliability coefficients of the other two criteria.)

Results

Univariate statistics

Descriptive statistics for the coloured and white apprentices on the four predictors and three criterion measures are given in Table 1.

Table 1 Comparisons of means and standard deviations of predictor and criterion measures for coloured and white apprentices

	White			Coloured			
Measure	*M*	*Sd*	*N*	*M*	*Sd*	*N*	*t* value
Interview	29,1	2,6	48	27,8	2,0	28	2,48[a]
Blox	32,7	4,9	52	34,7	6,8	51	1,78
HLFCT	17,0	4,2	48	18,3	3,7	28	1,26
Computation	14,7	3,5	52	15,1	3,8	51	0,53
Practical	73,8	7,2	52	72,0	8,4	51	1,16
Theory	81,1	11,4	52	77,6	11,8	51	1,53
Evaluation	72,2	6,2	52	72,1	9,4	51	0,07

[a] $P < 0,02$

As will be noted from the data presented in the table, a significant difference between the two groups was recorded for only one measure, namely, the panel interview ($t = 2,48$, $P < 0,05$). The mean scores recorded here suggest that on the average white apprentices were rated more positively than coloured apprentices. Although the difference is statistically significant, it is certainly not large in behavioural terms. Thus, generally speaking, the two groups compared well on both predictor and criterion measures.

Bivariate relationships

Naturally a major focus of this study was to determine how strong the predictors were related to the criteria of job performance of the apprentices. These relationships are expressed in terms of the correlations between the respective measures. These data are given in Tables 2, 3, and 4 for the practical criterion, theory criterion, and the supervisor's evaluation, respectively.

Table 2 Comparisons of correlations between predictors and practical criterion for white and coloured apprentices

	Practical scores		
Predictors	Whites	Coloureds	*t* values
Interviews	0,26	0,13	0,54
Blox	0,02	0,16	0,69
HLFCT	-0,03	0,48[b]	2,22[a]
Computation	-0,04	-0,12	0,40

[a] $P < 0,05$; [b] $P < 0,01$

Table 3 Comparisons of correlations between predictors and theory criterion for white and coloured apprentices

	Theory scores		
Predictors	Whites	Coloureds	*t* values
Interview	0,03	-0,11	0,56
Blox	-0,04	0,07	0,54
HLFCT	0,14	0,26	0,51
Computation	0,21	-0,01	1,10

Table 4 Comparisons of correlations between predictors and supervisor's evaluations for white and coloured apprentices

	Evaluations		
Predictors	Whites	Coloureds	*t* values
Interview	0,36[b]	-0,02	1,59
Blox	0,20	0,30[a]	0,53
HLFCT	0,05	0,23	0,74
Computation	0,15	0,22	0,36

[a] $P < 0,05$; [b] $P < 0,02$

A number of pertinent observations can be made from the figures provided in the three tables. First, it is obvious from the sizes of the correlations that the four tests have not shown strong associations with the three criteria. In fact, only three correlations are significantly different from zero. In Table 2 it is shown that the correlation between the HLFCT and the practical criterion for coloured apprentices of 0,48 is significant ($P < 0,01$) and in Table 4 the correlations between the panel interview and the supervisor's evaluation for whites is significant ($r = 0,36$, $P < 0,02$) as is the correlation between the Blox test and the supervisor's evaluation for coloureds ($r = 0,30$, $P < 0,05$). It is evident from the data in Table 3 that none of the predictors showed a significant relationship with the theory criterion. Thus, against this background it seems as if the tests have only limited utility in predicting future job-related performance.

A second important series of observations which can be made regarding the bivariate data presented in the last three tables concerns the differences between the correlations for the two groups of apprentices. As will be seen from the t values quoted in each table, a significant difference between the racial groups was recorded in only one case, namely, for the correlation between the HLFCT and the practical criterion. In this case the correlation of -0,03 for the white apprentices is significantly smaller than the correlation of 0,48 recorded for the coloured apprentices ($t = 2,22$, $P < 0,05$). Apart from this case it can be deduced that little evidence has been found for the presence of differential validity between white and coloured apprentices.

Evaluation of test fairness

In order to establish the degree of test fairness of the four selection measures, the procedure recommended by Lawshe (1983) was followed to convert predictor and criterion scores into scaled T values and, subsequently, calculating mean prediction errors. The means and standard deviations of the prediction errors for all the predictor/criterion pairs for white and coloured apprentices are given in Table 5.

Table 5 Comparisons of prediction error means and standard deviations for predictor/criterion pairs of white and coloured apprentices

Predictor/criterion pairs	Whites		Coloureds		
	M	*Sd*	*M*	*Sd*	*t* values
Interview/Practical	1,39	11,22	-2,52	11,79	1,44
" /Theory	1,07	14,01	-1,53	14,57	0,77
" /Evaluation	2,38	10,08	1,75	12,28	0,25
Blox/Practical	-2,48	11,28	2,87	13,26	2,21[a]
" /Theory	-2,64	12,74	3,65	14,22	2,37[b]
" /Evaluation	-1,71	9,66	1,77	13,01	1,54
HLFCT/Practical	-2,25	13,93	2,07	10,09	1,56
" /Theory	-2,58	13,81	3,07	13,13	1,75
" /Evaluation	-1,27	12,84	6,35	12,03	2,55[b]
Computation/Practical	-1,54	12,35	1,54	15,05	1,14
" /Theory	-1,70	12,13	2,32	14,59	1,52
" /Evaluation	-0,77	11,20	0,43	16,95	0,43

[a] $P < 0,03$; [b] $P < 0,01$

Test unfairness can be regarded as having occurred if a significant difference between the means for the respective prediction errors is recorded. A scrutiny of the figures in Table 5 reveals that in three out of the 12 cases statistically significant differences occurred between the white and the coloured mean scores, namely, for the Blox/practical ($t = 2,21$, $P < 0,03$), Blox/theory ($t = 2,37$, $P < 0,01$), and HLFCT/evaluation ($t = 2,55$, $P < 0,01$) pairs. On the basis of these data it can be concluded that the test underpredicted the criterion scores for a group of people, i.e., discriminated against that group. The question now arises against which group did these tests discriminate? A glance at the data in Table 5 reveals that in all three cases the means of the white apprentices are negative while those for coloured apprentices are positive. A positive mean indicates that the test has overpredicted the criterion score while a negative score implies underprediction of the criterion. Consequently, the results of this study suggest that in the three cases where unfairness occurred, the tests predicted unfairly for *white* apprentices. In operational terms this means that if cutting scores were used on the tests as part of a selection strategy, more white applicants with the potential for achieving job success would have been rejected than would be the case with coloured applicants.

Conclusion

The outcomes of this study are interesting from a number of points of view. First, it is evident that the utility of the selection instruments used by the organization in which this research was conducted is limited. The results of the study showed that the validity coefficients of the predictor/criterion relationships are generally very low. This means that unless the selection ratios which prevail in the relevant labour markets are particularly favourable, these instruments have little benefit as selection strategies.

A further interesting conclusion that can be drawn from the data obtained is the lack of evidence for differential validity of the respective tests for the two racial groups. It is interesting to note that much of the more recent research which has been conducted in the United States of America is also pointing in this direction, namely, that in rigorously designed studies it becomes less likely that a test will show different prediction patterns for different groups (Schmidt & Hunter, 1981). Coupled to this, the results recorded here also showed that there is limited evidence that the tests evaluated in this exercise acted in an unfair way toward a given racial group. In only three out of 12 comparisons were there indications that the tests were discriminating. Of particular interest is the finding that the Blox test and the HLFCT underpredicted the criterion scores for white apprentices and not for coloured apprentices as may

have been expected. In effect, this means that more white applicants with the potential of being successful job performers are likely to be screened out at time of hiring than coloured applicants. At first glance this finding appears contrary to the commonly held belief that psychological tests discriminate against culturally disadvantaged persons. It is, however, interesting to note that similar results have also been recorded in the United States of America. For example, Cascio (1987:174) concluded from a review of research that there is evidence that prediction systems 'slightly overpredicted' minority group performance.

The situations between South Africa and America are, of course, vastly different. On the one hand there is little reason to believe that a theoretical rationale exists for race as a moderator in test-criterion relationships in the American context. 'After all, blacks and whites live in the same society, watch the same television shows, attend similar schools, etc.' (Schmidt & Hunter, 1978:216). On the other hand, the South African society is characterized by inequalities in living standards and life styles with black people being particularly deprived of growth opportunities. Thus, there are good reasons to believe that race could act as a moderator when predicting job success with psychometric measures. Consequently, it is felt that much more research needs to be carried out in which the predictor-job performance patterns of blacks and whites are investigated. Taylor & Radford (1986:80) described the need for such research rather well when they noted: '...the political structures and the associated socio-economic differentiation, based on the forced ethnic segregation of the South African population, has played and continues to play an important role in the cognitive development of the members of the various groups [which]...should be reflected in psychometric test scores'.

References

Bleazard v Argus Printing & Publishing Co Ltd. (1983). *Industr. Law J.*, vol.4, (IC)60.

Cascio, W.F. 1987. *Applied Psychology in Personnel Management.* Englewood Cliffs, NJ: Prentice Hall.

Ehlers, D.E. 1982. Dispute settling and unfair labour practices. *Industr. Law J.*, vol.3, 11–21.

Golding, M. 1985. SAAWU in the Western Cape — The Continental China strike. *S. Afr. Labour Bull.*, vol.10, 57–74.

Lawshe, C.H. 1983. A simplified approach to the evaluation of fairness in employee selection procedures. *Personnel Psychol.*, vol.36, 601–608.

Mahlangu v CIM Deltak, Gallant v CIM Deltak (1986). *Industr. Law J.*, vol.7, (IC)346.

Raubenheimer, I. van W. 1983. Probleemareas by die gebruik van dieselfde of afsonderlike gestandardiseerde toetse vir verskillende bevolkingsgroepe. In J.F. Vorster (Ed.), *Simposium oor die problematiek wat ontstaan by die gebruik van dieselfde of afsonderlike toetse vir verskillende bevolkingsgroepe.* Pretoria: Human Sciences Research Council.

Schmidt, F.L. & Hunter, J.E. 1978. Moderator research and the law of small numbers. *Personnel Psychol*, vol. 31, 215–232.

Schmidt, F.L. & Hunter, J.E. 1981. Employment testing: Old theories and new research. *Am. Psychologist*, vol. 36, 1128–1137.

Taylor, J.M. & Radford, E.J. 1986. Psychometric testing as an unfair labour practice. *S. Afr. J. Psychol.*, vol.16, 79–86.