Journal of Industrial Psychology, 2001, 27(2), 32-38
Tydskrif vir Bedryfsielkunde, 2001, 27(2), 32-38

DETERMINING DIFFERENTIAL ITEM FUNCTIONING AND ITS EFFECT ON THE TEST SCORES OF SELECTED PIB INDEXES, USING ITEM RESPONSE THEORY TECHNIQUES

P SCHAAP
Department of Human Resources Management
University of Pretoria

ABSTRACT

The objective of this article is to present the results of an investigation into the item and test characteristics of two tests of the Potential Index Batteries (PIB) in terms of differential item functioning (DIF) and the effect thereof on test scores of different race groups. The English Vocabulary (Index 12) and Spelling Tests (Index 22) of the PIB were analysed for white, black and coloured South Africans. Item response theory (IRT) methods were used to identify items which function differentially for white, black and coloured race groups. The effects of the differences between the item characteristic curves (ICCs) of the three race groups on the test characteristic curve (TCCs) were studied. The items identified as biased (DIF) appeared to have a negligible effect on the test scores of Index 12 and Index 22 at the different ability levels for the groups considered. It can be concluded that the tests do not appear to discriminate unfairly, due to DIF, against race groups.

OPSOMMING

Die doel van hierdie artikel is om die resultate van 'n ondersoek na die item- en toetseienskappe van twee PIB (Potential Index Batteries) toetse in terme van itemsydigheid en die invloed wat dit op die toetstellings van rassegroepe het, weer te gee. Die Potential Index Batteries (PIB) se Engelse Woordeskat (Index 12) en Spellingtoetse (Index 22) is ten opsigte van blanke, swart en gekleurde Suid-Afrikaners ontleed. Itemresponsteorie (IRT) is gebruik om items te identifiseer wat as sydig (DIF) vir die onderskeie rassegroepe beskou kan word. Die effek van die verskille in itemkarakteristieke kurwes van drie groepe is op die toetskarakteristieke kurwe ondersoek. Dit blyk dat die items wat as DIF geïdentifiseer is, 'n onbenullige effek op die toetstellings van Index 12 en 22 vir die verskillende vermoënsvlakke van die drie groepe het. Dit kom voor dat die toetse nie as gevolg van DIF onbillik teen blank, swart of kleurlingrassegroepe diskrimineer nie.

The new Employment Equity Act (1998) places all test developers and users under an obligation to consider the impact of psychometric assessments on different groups as carefully as they consider other technical psychometric issues. The importance of the incorporation of this requirement in the design of psychometric instruments cannot be overemphasised. The fact that some tests may discriminate unfairly against certain groups has become a matter of primary concern in South Africa.

What complicates the issue of unfair discrimination is that differences in the experiential backgrounds of groups or individuals inevitably manifest themselves in test performance. Insofar as culture affects behaviour, cultural influences will and should be detected by such a measure. Sometimes, inexplicable differences in personality, cognition and factors involved in the test situation itself have an effect on the different performances on test items of different cultural groups (Scheuneman, 1985). However, if all cultural differentials which cause unfair discrimination related to differential item functioning (DIF) are ruled out from a test, the content validity of the test may be compromised. In an effort to include in tests only items common to different cultures or sub-cultures, content may be chosen that has limited scope in terms of the construct measured.

Insofar as the assumptions of the latent ability theory in item response models for unidimensional measures allow for differential item functioning, only one dominant component or primary trait influences test performance. This assumption, however, does not exclude the influence of secondary factors or traits on test performance. The secondary factors or traits that can have an impact on test performance in addition to the dominant component include cognitive, personality and test-taking factors. In terms of DIF, the apparent differences in the primary ability (when, in fact, there are no such differences) may be the consequence of secondary latent traits. Accordingly, DIF or item bias can be defined as the difference between two groups in the probability of an individual providing the correct response to an item, given the same primary or underlying ability. This means that, if an item is unbiased, the probabilities of correct responses at each ability level must be identical, apart from sampling error, across different populations of interest (Hambleton & Swaminathan, 1990).

However, the interpretation and evaluation of test scores are not based on the individuals' responses to items, but rather on scale scores and configurations of scores. Tests are most valuable if the test level rather than the individual item level is the basis of comparison. Some items might have lower content validity than other items in the test, and focusing on specific items might detract from the overall value of the test in assessing ability. Focusing on item level only without considering the cumulative effect on the total score would mean moving away from total information of the scale (Pope, Butcher & Seelen, 1994). Thus, the cumulative effect of DIF on the test score for the groups in question should be investigated before any conclusions can be reached on the level of unfair discrimination present in the test.

The objective of this article is to present the results of an investigation into the item characteristics of two tests of the Potential Index Batteries (PIB) in terms of DIF and the effect thereof on the test scores of different race groups.

METHOD

Strategy for identifying DIF

The three-parameter logistic item response theory model was used to identify DIF. The a, b and c parameters were obtained for each group by means of the marginal maximum likelihood (MML) procedure and the EM algorithm using the Xcalibre Item Parameter Estimation Programme. The programme's calculations include standardised residuals to indicate how well the response data fit the selected item response theory (IRT) model for the item parameters estimated. The statistical properties of the MML technique seem to have high levels of consistency. By implementing the MML technique, reasonable estimates of IRT item parameters can be derived from short tests (e.g. 25 items) and small samples of examinees (e.g. less than 1 000)(Mislevy & Bock, 1982).

Versoeke vir herdrukke moet gerig word aan: P. Schaap, Department of Human Resource Management, University of Pretoria, Pretoria 0002

TABLE 1
BIOGRAPHICAL INFORMATION ON THE SAMPLES

White		Black				Coloured	
Education:	Percentage	Education:		Percentage		Education:	Percentage
Standard 10	90%	Standard 10			88%	Standard 10	89%
Higher education	10%	Higher education			12%	Higher education	11%
First language:		First language:				First language:	
Afrikaans	50%	Zulu	8 %	Sesotho	10%	Afrikaans	62%
English	40%	Sepedi	15%	Tswana	21%	English	6%
Missing data	10%	Tsonga	5%	Xhosa	22%	Missing data	32%
		Venda	4%	SeSwati	1%		
		Missing data	14%				

Linn, Levine and Wardrop (1981) proposed a strategy using the area between the item characteristic curves for comparison and focal groups to determine bias. The strategy can be explained as follows: According to the three-parameter logistic model, the conditional probability Pi (⊖) that a person randomly chosen from all those with ability (⊖) will answer item i correctly, is a function of ⊖ and three item parameters. Each item is characterised by three item parameters: the item discrimination, a; the location or difficulty of the item, b; and the lower asymptote or probability that persons with extreme low ability will respond correctly to the item, c. The graph of Pi (⊖) as a function of⊖ is called the item characteristic curve (ICC) for item i. According to the model, the probability of getting the item right is completely determined by ⊖ and the three item parameters. More specifically, members of different groups with equal ability should have the same probability of correctly answering an item. In other words, the conditional probabilities, P (⊖), and their graphs should be invariant from one group to another if the item is not biased.

Since the item parameters have to be calibrated separately for each group, the item parameters were standardised on b*i* for each group before comparing the ICCs (Hambleton & Swaminathan, 1990). The c parameter was determined for the combined group and assumed to be fixed and equal for the different sub-groups which were compared.

Factor analysis by means of the SPSS statistical package was used to check whether the assumption of unidimensionality of the test items was reasonable. The phi correlation was used as a measure of the relationship between two dichotomous variables. It is commonly believed that using phi correlation leads to a factor solution with too many factors, some of them difficulty factors due to the range of item difficulties among the items in the test. A second order factor analysis was then done to determine the actual number of factors underlying the intercorrelation of the first order factors (Schepers, 1992).

The Mantel-Haenszel chi-square statistic, as calculated by means of the Biasx Programme of the HSRC, was used to cross-validate the results obtained on the IRT model (Holland & Thayer, 1988). Using multiple methods in studying DIF is generally recommended. The chi-square statistic can be considered a close approximation of the item response theory approach. The Mantel Haenszel statistics, in addition to the chi-square statistic, include the delta statistic as an indicator of the level and the direction of DIF. The calculations of Mantel Haenszel chi-square statistics were performed by the University of Pretoria's Network and Support Services.

Determining the effect of DIF on test scores

The Test Characteristic Curve (TCC) is the sum of P (⊖) for the items included in the test. The TCC is an estimation of the proportion of items answered correctly at each ⊖ level (Hambleton & Swaminathan, 1990). Therefore, the difference in the TCCs at each ability level for different groups provides an indication of the difference in the standardised ICC for each group. The cumulative effect of DIF on the test score can be estimated through calculating the net differences in ICCs of biased items for the groups in question.

Data analysed

Data for the analyses reported below are based on the English Vocabulary (Index 12) and English Spelling Ability (Index 22) tests of the PIB, consisting of 20 and 25 items respectively (Erasmus & Minnaar, 1997). Index 12 requires the testee to indicate which one of five alternative words has more or less the same meaning as a specified word. Index 22 requires the testee to indicate the correct spelling of a specific word, given five alternatives. Both measures were developed for post-matriculants. No time limits were imposed for the completion of the tests. The test chi-square and item response data were obtained from job applicants. A convenience sample consisting of 609 white and 694 coloured candidates was used. The sample represents the total number of available records for the particular groups. A random sample of 677 black candidates, randomly drawn from the 5 000 available data records for black candidates, was used. Frequency distributions indicating the first language and educational characteristics of each of the samples are set out in Table 1. For the purposes of this study, the black and coloured race groups were considered to be the previously disadvantaged groups and accordingly treated as the focal group. The white group was consequently treated as the comparison group. In this respect, the white-black comparisons and the white-coloured comparisons were considered the main areas of interest relating to test fairness.

Indices of DIF

Three indices of DIF involving areas between ICCs were computed (Linn et al., 1981). The three bias indices used for the results reported below are the following:

1. Base high area (BHA): the area, if any, between the ICCs for the groups compared where the ICC for the focal group is above that of the ICC for the comparison group.
2. Base low area (BLA): the area, if any, between the ICCs for the groups compared where the ICC for the focal group is below that of the ICC for the comparison group.
3. Square root of the sum of squares: the square root of the sum of the squared differences between ICCs in the region of ⊖ = -3 to ⊖ = +3.

An item with a large BHA but small or zero BLA would be considered to be DIF against the comparison group. The direction of DIF would be just the opposite for an item with a large BLA but zero or small BHA. The bias in an item with large BHA and large BLA would depend upon the distribution of ability in the contrasted groups of examinees. The square root of the sum of squares provides an index of total DIF in the region of ⊖ = -3 to ⊖ = ⊖+3. Linn et al. (1981) used a 0,2 cut-off as an indication of possible DIF. Although it cannot be claimed that the value of 0,20 corresponds to a significance statistic of 0,10 or 0,05, it should be a good approximation thereof (Hulin, Drasgow & Parsons, 1983) and indicates a high possibility of DIF.

Due to the sample size sensitiveness of the MH chi-square statistic, a cut-off of 10,83 (at a significance level of 0,01) as an indication of DIF was applied (Raju, Drasgow & Slinde, 1993; Raju, 1990).

RESULTS

The test score summary statistics for black, white and coloured candidates are set out in Table 2. The raw score means for the black and coloured groups on Index 12 are approximately the same, whereas the mean for the white group is approximately 3 points or 0,70 standard deviation greater than the mean for the black and coloured groups. The mean for the black group on Index 22 is approximately 2 points or 0,46 standard deviation greater than that of the white group and 1 point or 0,34 standard deviation greater than that of the coloured group. Each of the above standard deviation comparisons made was based on the standard deviation values of the group with the greatest mean value for the test. The difference between the mean values for each of the comparisons made are statistically significant ($p \leq 0,01$), except for the comparison between the black and coloured groups on Index 12.

TABLE 2
SUMMARY STATISTICS FOR WHITE, BLACK AND COLOURED GROUPS

		White (N=609)	Black (N=677)	Coloured (N=694)
Index 12	Mean	12,96	10,03	10,00
	SD	4,16	3,40	4,31
Index 22	Mean	11,96	13,79	12,46
	SD	4,35	3,96	4,04

Unidimensionality

The principal factor analyses using phi coefficients based on the total group of 1 980 respondents provided evidence of the unidimensionality of Index 12 and Index 22. Index 12 yielded four eigenvalues greater than unity, with the highest eigenvalue of 3,4 accounting for 18% of the total variance. The second and third eigenvalues accounted for substantially smaller percentages of the total variance (7,1 % and 3,1 % respectively). According to the scree test, there appeared to be a single dominant factor. A second order factor analysis was done which yielded one eigenvalue greater than unity. This indicates one dominant factor, with an eigenvalue of 2,13 accounting for 53 % of the total variance. The same procedure as above was followed with Index 22. The first order factor analysis yielded eight eigenvalues greater than unity, with the highest eigenvalue of 3,24 accounting for 13 % of the total variance. The second and third eigenvalues accounted for substantially smaller percentages of the total variance (5,1 % and 4,7 % respectively). The scree test indicated a single dominant factor. The second order factor analysis yielded one eigenvalue greater than unity which indicates the presence of one dominant factor, with a value of 2,10 accounting for 26 % of the total variance. According to Hulin et al. (1983), IRT models can be applied to moderately heterogeneous item sets.

Excaliber calibrations

The item parameters estimates for the white, black and coloured groups were determined. The calculated standardised residuals indicate that all the response data fit the 3 Parameter IRT model for the item parameters estimated by means of the Excaliber Calibrations Programme. The item parameters were standardised on parameter b for the black and coloured groups (focal groups), using the mean b parameter of the white group (comparison group). The c parameter was kept constant for each of the groups, based on the c parameter for the total group (Raju, Drasgow & Slinde, 1993).

DIF indices: Index 12 and 22

The DIF statistics for the white-black and white-coloured comparisons are set out in Tables 3 and 4. These tables show, for each item, the base high area (BHA), base low area (BLA), square root of the sum of squares (DIF), the Mantel-Haenszel (MH) chi-square and delta statistic. DIF and MH chi-square values that were considered to be significant are marked with an asterisk (*). High values on BHA and low values on BLA indicate DIF in favour of the black and coloured groups while high values on BLA and low values on BHA indicate DIF in favour of the white group. A negative value on the MH delta statistic indicates DIF in favour of the black and coloured groups and a positive value indicates DIF in favour of the white group.

In the case of the white-black group comparison, nine biased items were identified based on the IRT method and ten items

TABLE 3
ITEM PARAMETERS AND DIF FOR INDEX 12 (VOCABULARY TEST)

Item	White (N=609)		Black (N=677)							Coloured (N=694)						Combined	
	a	b	a	b	BLA	BHA	DIF	Chiq	Delta	a	b	BLA	BHA	DIF	Chisq	Delta	c
1	0,74	0,12	1,09	-0,27	0,03	0,32	0,18	0,0	0,0	1,05	-0,10	-0,10	0,36	0,13	1,0	0,3	0,24
2	0,97	0,81	1,07	0,74	0,00	0,01	0,04	1,7	-0,4	1,08	0,87	0,00	0,07	0,03	0,4	-0,2	0,27
3	0,97	-1,16	1,01	-0,83	0,24	0,00	0,13	3,4	0,7	0,99	-1,46	0,00	0,19	0,11	5,5	-1,0	0,24
4	0,87	-0,00	1,09	-1,28	0,00	0,97	0,50*	107,4*	-3,7	1,15	-0,03	0,07	0,10	0,08	0,0	0,1	0,23
5	0,78	0,79	1,56	0,82	0,14	0,09	0,10	6,7	0,8	0,91	0,69	0,00	0,10	0,05	1,2	-0,3	0,26
6	1,09	-1,21	0,96	-0,56	0,49	0,00	0,27*	19,6*	1,6	1,26	-0,13	0,83	0,00	0,47*	118,5*	4,1	0,22
7	0,91	-0,76	1,14	-0,51	0,00	0,57	0,31*	43,5*	-2,5	1,22	-0,62	0,14	0,04	0,10	2,8	0,6	0,23
8	0,79	-1,37	0,93	-0,91	0,31	0,00	0,17	9,6	1,1	1,02	-1,58	0,01	0,19	0,11	5,1	-1,0	0,24
9	0,87	-0,20	0,93	-0,74	0,00	0,39	0,20*	20,5*	-1,5	0,87	-0,48	0,00	0,20	0,10	5,8	-0,8	0,27
10	0,86	0,33	1,17	1,18	0,66	0,00	0,36*	44,3*	2,1	1,02	0,28	0,03	0,07	0,05	0,0	-0,1	0,22
11	1,14	-0,93	1,22	0,12	0,80	0,00	0,45*	79,0*	3,0	1,12	-0,99	0,00	0,05	0,03	1,6	-0,5	0,23
12	1,04	0,06	1,20	1,18	0,89	0,00	0,49*	57,6*	2,5	1,22	0,77	0,57	0,00	0,32*	18,9*	1,4	0,20
13	1,03	1,72	1,20	0,57	0,00	0,91	0,52*	51,5*	-2,7	1,24	1,37	0,00	0,26	0,17	0,2	0,2	0,16
14	0,82	0,37	0,96	-1,04	0,00	1,16	0,49*	39,3*	-2,0	0,85	0,08	0,00	0,20	0,10	0,2	0,1	0,28
15	0,99	0,18	1,27	0,21	0,08	0,06	0,07	0,50	0,3	1,20	0,21	0,00	0,09	0,06	0,0	0,0	0,19
16	0,83	-0,91	0,93	-0,44	0,33	0,00	0,18	17,9*	1,4	0,86	-0,89	0,00	0,00	0,12	0,1	0,1	0,26
17	1,10	-0,38	1,22	-0,23	0,12	0,00	0,08	1,7	0,5	1,10	-0,23	0,12	0,00	0,07	0,0	0,1	0,21
18	0,86	-0,44	1,04	-0,17	0,22	0,00	0,12	7,5	0,9	1,23	-0,17	0,24	0,04	0,15	7,9	0,9	0,22
19	0,98	0,10	1,11	0,17	0,07	0,01	0,04	0,8	0,3	1,02	0,09	0,03	0,03	0,03	0,2	-0,1	0,22
20	0,85	-0,96	1,07	-0,81	0,12	0,03	0,08	0,6	0,3	0,93	-1,42	0,00	0,33	0,18	14,2*	-1,4	0,24

* : DIF $\geq$ 0.,2 is probably statistically significant
* : DIF $\geq$ Chisq 10,88 is taken as significant
BHA : Base high area
BLA : Base low area
DIF : Square root of the sum of squares
Chisq : Mantel-Haenszel chi-square statistic
Delta : Mantel-Haenszel delta statistic
a : IRT a-parameter
b : IRT b-parameter
c : IRT combined c-parameter for total group

based on the MH technique. Of the ten items identified by means of the MH technique, nine items were also identified as biased using the IRT method. Only Item 16 was not included by the IRT method. The results indicate a very high similarity between the IRT and MH technique in terms of the identification of DIF. Raju, et al. (1993) and Raju (1990) reported a similarly large percentage of overlap with significant signed and unsigned areas between two item response functions and the MH technique using more or less the same cut-off on the MH chi-square statistic (i.e. 9 and 10,88 respectively). In both studies, the MH technique identified slightly more biased items than the IRT technique (i.e. overlap of 0,80). These findings provide evidence of the accuracy of the square root of the sum of squares as a DIF index using a cut-off of 0,20 in the current study.

It is interesting to note that five items, i.e. items 4, 7, 9, 13 and 14 seemed to favour the black group, while four items, i.e. items 6, 10, 11 and 12 (using the IRT technique), and five items, including item 16 (using the MH technique), favoured the white group. The ICCs for Items 6 and 13 are illustrated as examples in Figures 1 and 2 respectively. Although the ICCs were substantially different for black and white groups for approximately half of the items, the direction of DIF did not consistently favour any of the groups. Five of the items that were identified as biased favoured the black group and the four remaining items favoured the white group. The estimated net effect of the nine items identified as biased on the difference between groups in the proportion of items answered correctly (TCC) is illustrated in Figure 3. The calculation of the estimated difference between groups on the TCC is based on the assumption that the items which were not identified as significantly biased had a difference in P (⊖) values of zero. As Figure 3 illustrates, the effect of the items identified as biased on the estimated difference in test scores between the white and black groups seems to favour the black group, but the effect can be considered to be minor. The largest difference occurs between ability level -1,50 and -0,50, but the estimated difference in test scores does not exceed 0,38 at any point. An average difference in the estimated test score of 0,19 occurred over the whole spectrum of abilities between -3 and +3, due to DIF. Thus, eliminating all DIF would have only a negligible effect on the group differences in the test as a whole, at the risk of reducing the validity of the test.

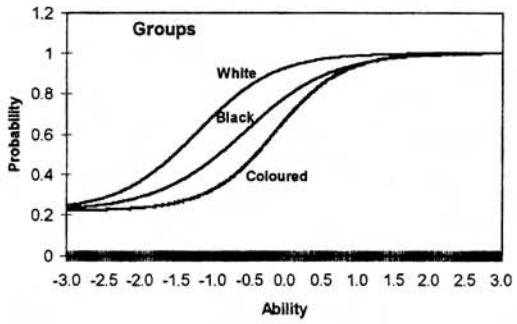

Figure 1: ICC, Item 6 (Index 12)

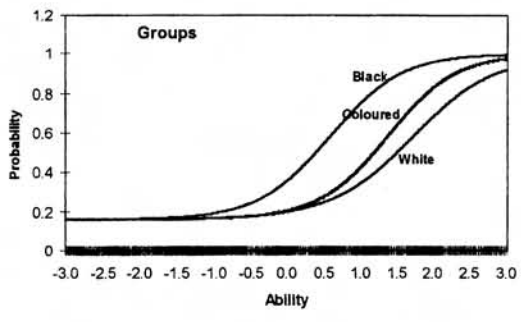

Figure 2: ICC, Item 13 (Index 12)

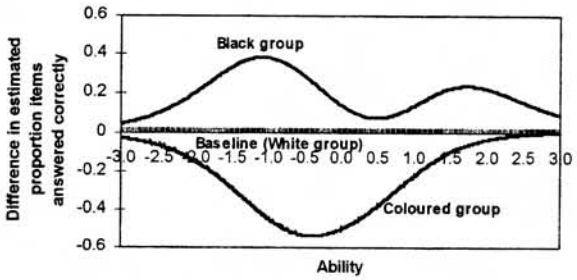

Figure 3: Difference in TCCs due to DIF items (Index 12)

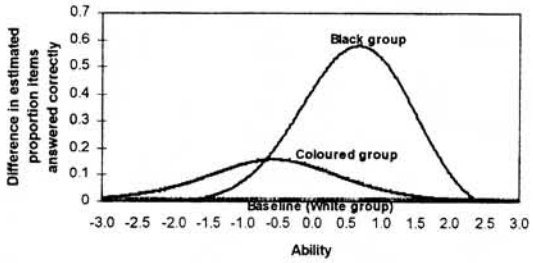

Figure 4: Difference in TCCs due to DIF items (Index 12)

Only two biased items were identified in the comparison of the white-coloured groups (using both the IRT and MH techniques) on Index 12. These items include items 6 and 12 and both favour the white group. The estimated effect thereof on the total test score can be considered to be very small and negligible (Figure 3), assuming that the difference between the groups on all the other non-biased items in the test is zero. The largest difference occurs between ability levels -0,5 and 0,00, but the estimated difference in the test scores does not exceed 0,54 at any point. An average difference in the estimated test score of 0,23 occurred over the whole spectrum of abilities between -3 and +3, due to DIF.

Index 22 was analysed using the same procedures as above. Nine biased items were identified. As with Index 12, there was a very large overlap between results obtained using the IRT and MH techniques to identify DIF. In the case of the white-black group comparison, nine biased items were identified based on the IRT method and eight items were found based on the MH technique. Of the nine items identified by means of the IRT technique, eight items were also identified as biased using the MH technique. Only item 24 was not included in the MH technique. As with Index 12, it is interesting to note that six items, i.e. items 5, 8, 11, 13, 17 and 22 seemed to favour, the black group. Three items, i.e. items 6, 10 and 24 (using the IRT technique), and two items, excluding item 24 (using the MH technique), favoured the white group. The estimated net effect on the test scores is illustrated in Figure 4 and can be considered to be very small and negligible. The largest difference in favour of the black group occurred between ability levels 0,50 and 1,00, but the estimated difference in the test scores did not exceed 0,58 at any point. An average difference in the estimated test score of 0,21 occurred over the whole spectrum of abilities between -3 and +3 due to DIF.

In terms of the white-coloured comparison, only one item was identified as biased, favouring the coloured group. As with the previous findings, the estimated net effect of DIF on the total score is negligible. The largest difference occurred between ability levels -1,00 and 0,00, but the estimated difference in the test scores did not exceed 0,16 at any point. An average difference in the estimated test score of 0,07 occurred over the whole spectrum of abilities between -3 and +3, due to DIF.

It is interesting to note that all the items identified as biased for each of the tests differed on each (ability) level without overlapping at any point. This signifies consistent DIF in favour of a particular group at all levels.

TABLE 4
ITEM PARAMETERS AND DIF FOR INDEX 22 (SPELLING TEST)

Item	White (N=609)		Black (N=677)							Coloured (N=769)							Combined
	a	b	a	b	BLA	BHA	DIF	Chisq	Delta	a	b	BLA	BHA	DIF	Chisq	Delta	c
1	0,96	-1,61	0,90	-1,31	0,45	0,67	0,12	4,7	1,1	0,92	-1,60	0,04	0,07	0,01	0,1	-0,2	0,23
2	0,97	-1,64	0,90	-1,49	0,11	0,00	0,06	1,7	0,7	0,99	-1,83	0,00	0,13	0,07	3,7	-1,0	0,23
3	1,03	-1,25	0,96	-1,60	0,00	0,24	0,14	2,4	-0,8	1,07	-1,16	0,06	0,00	0,04	0,0	0,0	0,23
4	0,94	-0,27	0,93	-0,62	0,00	0,25	0,14	2,6	-0,6	0,93	-0,80	0,00	0,39	0,21*	21,0*	-1,5	0,23
5	0,99	1,09	0,74	0,32	0,00	0,57	0,29*	41,6*	-1,9	1,06	1,17	0,00	0,06	0,04	4,3	0,6	0,25
6	0,84	-0,93	0,89	0,01	0,67	0,00	0,34*	47,0*	2,4	0,94	-0,74	0,12	0,00	0,07	0,4	0,2	0,24
7	0,90	0,02	0,78	0,55	0,37	0,00	0,19	6,7	0,8	0,87	0,19	0,24	0,00	0,06	1,9	0,4	0,26
8	0,88	-0,06	0,86	0,48	0,00	0,39	0,20*	15,3*	-1,2	0,98	0,21	0,24	0,00	0,10	2,9	0,5	0,25
9	1,02	1,08	0,86	0,73	0,00	0,28	0,15	8,8	-0,9	0,99	1,39	0,24	0,00	0,13	1,5	0,4	0,20
10	0,93	0,69	0,84	1,32	0,44	0,00	0,24*	13,7*	1,1	0,88	0,53	0,00	0,12	0,06	0,2	-0,2	0,24
11	0,97	1,89	0,85	0,55	0,00	0,96	0,50*	54,6*	-2,2	0,90	1,77	0,00	0,09	0,05	3,3	-0,5	0,24
12	1,01	0,83	0,91	1,06	0,17	0,00	0,09	2,3	0,5	0,97	1,03	0,15	0,00	0,08	1,1	0,3	0,23
13	1,04	1,19	0,86	0,56	0,00	0,48	0,25*	17,1*	-1,3	0,97	1,4	0,17	0,00	0,10	0,0	0,0	0,23
14	1,03	1,32	0,88	1,67	0,23	0,00	0,15	0,7	0,3	1,02	1,13	0,00	0,14	0,08	4,0	-0,6	0,21
15	1,00	1,64	0,88	2,10	0,30	0,00	0,19	4,9	0,7	1,01	1,77	0,09	0,00	0,05	0,9	0,3	0,19
16	0,91	1,58	0,88	1,80	0,14	0,00	0,08	0,8	0,3	0,93	1,54	0,00	0,02	0,01	0,0	-0,1	0,26
17	0,90	0,57	0,83	-0,09	0,00	0,49	0,24*	22,0*	-1,4	0,83	0,30	0,00	0,19	0,10	7,4	-0,8	0,25
18	1,00	2,21	0,90	2,18	0,00	0,01	0,03	5,2	-0,7	1,02	1,92	0,00	0,18	0,12	3,8	-0,6	0,19
19	1,00	1,81	0,91	2,14	0,20	0,00	0,12	0,2	0,2	0,99	1,82	0,01	0,00	0,01	0,9	-0,3	0,24
20	1,02	2,44	0,93	2,56	0,04	0,00	0,04	0,0	0,1	1,02	2,20	0,00	0,14	0,09	0,0	0,0	0,20
21	0,96	1,92	0,91	1,87	0,00	0,04	0,11	2,4	0,5	1,02	1,64	0,00	0,18	0,11	0,20	-0,2	0,23
22	0,94	1,21	0,73	0,57	0,00	0,46	0,23*	13,3*	-1,1	0,99	1,56	0,24	0,00	0,13	5,9	0,7	0,28
23	0,91	1,56	0,83	1,29	0,00	0,19	0,10	0,3	-0,2	0,95	1,36	0,00	0,13	0,07	0,0	0,0	0,24
24	1,08	1,58	0,95	2,09	0,34	0,00	0,22*	3,7	0,7	1,03	1,83	0,18	0,00	0,11	1,1	-0,4	0,18
25	0,95	0,97	0,86	1,07	0,08	0,01	0,05	0,9	0,3	1,10	1,02	0,18	0,14	0,05	4,5	0,6	0,23

* : DIF $\geq$ 0,2 is probably statistically significant
* : DIF $\geq$ Chisq 10,88 is taken as significant Delta

Base high area
BLA : Base low area
DIF : Square root of the sum of squares

Chisq : Mantel-Haenszel chi-square statistic
: Mantel-Haenszel delta statistic

a : IRT a-parameter
b : IRT b-parameter
c : IRT combined c-parameter for total group

DISCUSSION

The reality of DIF is a phenomenon that must be acknowledged and appropriately dealt with in tests designed for heterogeneous groups. However, the exclusion of all DIF from tests with the objective of developing valid and culture-free tests may be an impractical solution. Based on the results of this study, it is clear that test items included in a test could be non-biased for specific groups but will not necessarily be non-biased for all groups. In accordance with the findings of this investigation, it is suggested that DIF need not be a limitation to ensuring cultural fairness, provided that the DIF does not cause a recognisable difference in the total test scores of different groups. Through the application of IRT methodology, it became clear that the biased items that were identified in the English Vocabulary and English Spelling tests did not cause a recognisable difference in test scores for the groups considered. Thus, the results from this investigation provide evidence of the overall cultural fairness of the PIB's English Vocabulary and Spelling tests for black and coloured race groups (focal groups) compared to white race groups (comparison group), irrespective of the DIF present. It must be noted that on both tests, considerably more items were identified as biased in the white-black comparisons, than in the white-coloured comparisons. Irrespective of this fact, it appears that for both the white-black comparisons and the white-coloured comparisons, DIF did not cause a recognisable difference in test scores. Therefore, it can be concluded that the number of items identified as biased is not necessarily related to specific levels of unfair discrimination in terms of test scores, but more specifically, the extent to which DIF consistently favours one group above the other group.

REFERENCES

Republic of South Africa (1998). Employment Equity Act. Pretoria: Government Printer.

Erasmus, P.F. & Minnaar, G.G. (1997). *Manual: Potential Index Battery (PIB) and The Advanced Potential Index Battery (Ad-PIB).* Pretoria: Potential Index Associates.cc.

Hambleton, R.K. & Swaminathan, H. (1990). *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In: H. Wainer & H.I. Braun (Eds.). *Test Validity.* Hillsdale NJ: Erlbaum.

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item response theory: application to psychological measurement.* Illinois: Dow Jones-Irwin.

Linn, R.L., Levine, C.N.H. & Wardrop, L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement,* 5(2), 159-173.

Mislevy & Bock, R.D. (1982). Implementation of an EM algorithm in the estimation of item parameters. In D.J. Weis (Ed.). *Proceedings of the 1982 item response theory and computerized adaptive testing conference.* Minneapolis: Psychometrics Methods Program, University of Minnesota.

Pope, H.S., Butcher, J.N. & Seelen, J. (1994). *The MMPI, MMPI-2 and MMPI-A in court: A practical guide for expert witnesses and attorneys.* Washington, DC: American Psychological Association.

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement,* 14, 197-207.

Raju, S.R., Drasgow, F. & Slinde, J.A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Education and Psychological Measurement,* 53, 301-313.

Schepers, J.M. (1992). *Toetskonstruksie: Teorie en praktyk.* Johannesburg: Randse Afrikaanse Universiteit.

Scheuneman, J.D. (1985). *Exploration of causes of bias in test items* (Research Report, 85-42). Princeton, NJ: Educational Testing Services.

User's Manual for the Xcalibre Marginal Maximum-Likelihood Estimation Program. (1995). St Paul:Assessment Systems Corporation.