Offprint from: European Journal of Psychological Assessment, Vol. 18, Issue 1, pp. 9–15

Effects of Ability Scale Purification on the Identification of dif

María J. Navas-Ara[1] and Juana Gómez-Benito[2]

[1]Universidad Nacional de Educación a Distancia, Spain, [2]Universitat de Barcelona, Spain

Keywords: Differential item functioning, ability purification, item response theory, restricted factor analysis, Mantel Haenszel, logit model, logistic regression

Summary: Research related to the detection of item bias or differential item functioning (dif) has proliferated in psychometric and applied psychological literature over the last 25 years. In fact, debate has been heated on the nature and, more particularly, on the methods for bias/dif detection. Today, conditional methods have obtained wide acceptance. However, these methods present a problem of circularity: If the test contains biased items, then a biased measure of the matching variable will be used for investigating dif. Thus, we investigate dif with a biased measure. The aim of this paper is to investigate the feasibility of improving item bias identification through "purification" of the ability measure used. Several bias-detection techniques are analyzed: Mantel-Haenszel statistic, logit model, logistic regression procedure, restricted factor analysis, and two item response theory (IRT)-based indices. Results show that purifying the ability scale improves item bias detection greatly, providing rates of correct identification close to 100% with all these techniques. IRT-based indices showed the greatest improvement.

Bias, or the currently preferred term differential item functioning (dif), is a long-standing topic of intense debate in psychometric literature. In fact, the issue of bias in testing has been a source of recurring, at times impassioned, social controversy throughout the history of mental measurement. Debate has been heated on the nature and, more particularly, on the methods for bias/dif detection. In fact, research related to the detection of item bias has proliferated in psychometric and applied psychological literature over the last 25 years.

According to Cole and Moss (1989), much of the work in dif since 1982 has focused upon χ^2 and item response theory (IRT)-based methods (for a review, see, e. g., Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993; Scheuneman & Bleistein, 1989). Unlike classical test theory-based methods, these two approaches are conditional methods. That is, dif is based on *differences* in item difficulty between groups given the level of ability (Mellenbergh, 1989).

However, conditional methods for dif detection share a common problem: These techniques contain a certain circular logic since the items being studied help to define the matching variable (observed test score or θ estimate). If the test contains biased items, then a biased measure of the matching variable will be used for investigating dif. Several procedures have been proposed to try to overcome this dilemma.

Among the non-IRT procedures are:

1) the iterative logit method (van der Flier, Mellenbergh, Ader, & Wijn, 1984; Kok, Mellenbergh, & van der Flier, 1985), in which the test is iteratively freed from biased items and all items are tested in the final step using an unbiased reduced test as ability indicator;
2) the two-stage version of the Mantel-Haenszel statistic, first suggested by Holland and Thayer (1988) and later used and evaluated by Clauser, Mazor, and Hambleton (1993), which consists of a purification of the matching criterion, eliminating in the second stage the items revealed as biased in the first stage.

Among the IRT procedures, Marco (1977) and Lord (1980) suggested a multistage estimation procedure that has yet to be used in practice. Park (1988) and Park and Lautenschlager (1990) have proposed a modification of this procedure, known as the modified-Lord test purification (M-LTP) method or iterative linking and ability scale purification (ILAP) method (Lautenschlager, Flaherty, & Park, 1994). As the literature reveals (Candell & Drasgow, 1988; Candell & Hulin, 1986; Drasgow, 1987; Hulin & Mayer, 1986; Kim & Cohen, 1992; Park & Lautenschlager, 1990; Segall, 1983), a simpler and equally effective procedure involves iterative procedures focused on equating or linking of groups. It has been repeatedly shown (e. g., Lautenschlager & Park, 1988) that item parameter linking methods are adversely affected by the presence of biased items and often produce inadequate linking results. In short, in this procedure equating is done by removing items identified as biased in the former iteration. Dif indices are recalculated in each iteration, but the estimation of item and ability parameters is carried out only once.

The present work takes into account these considerations. Specifically, the aim of this paper is to investigate the feasibility of improving dif identification through the "purification" of the ability measure used. The effect of ability scale purification on identification of dif is investigated through the use of six different dif detection techniques: Mantel Haenszel, logit model, logistic regression, restricted factor analysis and two IRT-based indices. In order to do this, simulated datasets are used.

Method

Data Generation

Item responses were generated using the one-parameter logistic item response model. In this model, the probability of an examinee answering correctly the *i*th item, given unidimensional latent trait θ, is given as:

$$P_i(\theta) = \frac{1}{1+\exp(-1.7(\theta - b_i)} \qquad (1)$$

where b_i is the difficulty parameter of item *i*. *b*s were sampled from a normal distribution N(0,1) and sample size, that is, test length, was 25. Table 1 contains these item parameters. θs were also sampled from a normal distribution N(0,1), and sample size was 1000. Therefore, the dimension of the matrix to be simulated was 1000 × 25, the responses of 1000 examinees to 25 test items.

Dichotomous item responses were generated by computing the probability of a correct response for each item and each examinee using equation (1). The probability of a correct response was then compared to a number sampled from a uniform distribution U(0,1). If the probability was less than the random number, the item response was scored as 0; otherwise, the item response was scored as 1.

Two matrices were generated following this procedure, one for the responses of examinees in the so-called reference group (R) and another for examinees in the focal group (F). The dimension of the two matrices was always the same: 1000 × 25.

The latent trait distribution was exactly the same for the reference and focal groups. That is, the same 1000 θ values sampled from the normal distribution were introduced into equation (1) for both groups.

In the test, 40% of the items were made to be biased. Biased items were randomly selected using a systematic sampling design with sampling fraction equal to 3, starting randomly at item number 2. Items were biased by decreasing their difficulty parameters by 0.75 prior to generating item responses for the focal group ($b_F = b_R - 0.75$ for 10 items), that is, prior to computing the probability of a correct response for each item with equation (1). This approach, known as unidimensional dif construction, has been followed in several dif studies (e. g., Kim & Cohen, 1992; Lim & Drasgow, 1990; Miller & Oshima, 1992; Rudner, Getson, & Knight, 1980; Shep-

Table 1. Item parameters used to generate response data*.

Item no.	*b*
1	*0.847*
2	*-0.051*
3	-1.060
4	*-1.160*
5	*-0.558*
6	0.440
7	1.311
8	*0.198*
9	-0.278
10	0.928
11	*2.287*
12	-0.250
13	0.242
14	*-1.231*
15	-1.086
16	1.177
17	*0.787*
18	-0.476
19	-0.900
20	*0.094*
21	-1.018
22	0.936
23	*1.217*
24	-2.275
25	0.513

*Items in *italics* are items made to be biased.

ard, Camilli, & Williams, 1985). Hence, dif is introduced in only one direction, so dif is uniform, that is, there is no interaction between ability level and group membership.

In summary, the same procedure was followed to generate the matrix for the reference and focal groups, the only difference being that when simulating the focal matrix, the difficulty values of 10 items were modified (items in italics in Table 1).

Three replications were made for the reference and focal groups in order to avoid capitalizing on chance. Thus, eight matrices in total were finally simulated.

A recovery analysis was carried out to determine the extent to which the generating parameters were captured in the simulated matrices. Before assessing the adequacy of parameters recovery, the mean and sigma method was used to transform parameter estimates to the underlying metric. As expected, all correlations between estimates and underlying parameters showed values greater than 0.9, and the root mean squared difference of true and estimated parameters varied between 0.4 and 0.6. Based on these results, recapture of the underlying item and ability parameters appeared to be reasonably good.

Dif Detection Methods

The effect of ability scale purification on the identification of dif was investigated using six different dif detection methods, some of which were based on Observed Conditional Invariance (OCI) models and others on Unobserved Conditional Invariance (UCI) models (Millsap & Everson, 1993).

OCI methods compare the item performance of comparable examinees in the reference and focal groups by using as matching criterion an observed measure of ability, most often, total test score. The methods herein considered are the Mantel-Haenszel technique (Holland & Thayer, 1988), the logit model (van der Flier et al., 1984), and the logistic regression procedure (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Analyses have been carried out with statistical packages BMDP for Mantel Haenszel and SPSS for logit model and logistic regression procedures. Readers are referred to the cited literature for details relative to these well-known techniques.

Unlike OCI methods, the matching criterion used by UCI methods is unobserved. The methods herein considered are two IRT-based indices and restricted factor analysis.

The *IRT-based indices* chosen to investigate dif in this study are measures of the area between item characteristic curves estimated in the reference and focal groups. In particular, two statistical indices were used to study dif: the signed area (SA) statistic as defined by Raju (1988, 1990), and the sum of squares 1 (SOS1) statistic (Shepard, Camilli, & Williams, 1984).

Two procedures have been used to help interpret the results obtained when calculating the first index. First, a significance test for the signed area was calculated for every item (Raju, 1990). Second, SA statistics were calculated in new datasets generated such that they could provide an adequate baseline for evaluating the results obtained when computing SA indices in the original datasets. This second procedure was also used with SOS1 indices.

The computer program used for estimating item and ability parameters was RASCAL (Assessment Systems Corporation, 1988). The procedure used to put estimates on the same scale prior to calculating SA and SOS1 indices was the mean and sigma method (Lord, 1980).

Restricted factor analysis is a seldom-used but fruitful technique for studying dif. Oort (1992, 1993) has proposed using it for investigating dif with respect to the potential violator *group membership*. In particular, according to his theory of potential violators, the score on item i of a randomly selected examinee may be modeled as:

$$u_i = m_i + a_i T + b_i G + D_i \quad (2)$$

where
m_i is the intercept parameter;
T is a latent variable representing true score on the trait of interest (total test score in the present work);
a_i is the population regression coefficient of item i on T;
G is the potential violator *Group membership*;
b_i is the population regression coefficient of item i on G, and
D_i is a residual factor containing error and true score on what is specific to item i (Oort, 1992).

An item is said to be biased with respect to a potential violator G if there is a direct effect of this violator on item i, that is, if $b_i \neq 0$. So the null hypothesis to be tested for every item in the test is H_0: $b_i = 0$.

Analyses were carried out with the program LISREL 8 (Jöreskog & Sörbom, 1993a) using weighted least squares estimation. PRELIS 2 (Jöreskog & Sörbom, 1993b) was previously used to obtain interitem tetrachoric correlations.

Criterion Purification Methods

A two-stage procedure was used for purifying the explicit criterion used for matching examinees of the same ability in the reference and focal groups with the Mantel Haenszel technique and the two IRT-based procedures,

Table 2. Dif detection rates using several techniques.

Technique	Iteration	Correct identification	False positive	False negative
Mantel Haenszel	First	.91	.15	–
	Second	1.00	–	–
Logit model	First	.73	.43	.02
	Last	1.00	–	–
Logistic regression	First	.75	.22	.30
	Second	.93	.02	.15
Restricted factor analysis	First	.71	.36	.18
	Last	.96	.07	–
Signed area (significance test)	First	.67	.48	.10
	Second	.97	.04	.02
Signed area (baseline)	First	.70	.48	.02
	Second	.99	.02	–
Sum of squares	First	.62	.57	.10
	Second	.97	.04	.02

and the ability indicator in logistic regression. The procedure is as follows.

The criterion is refined by eliminating biased items based on a preliminary dif analysis, so the purified criterion/ability indicator is composed exclusively of items revealed as unbiased by the previous dif analysis, and a new dif analysis is conducted using the refined measure.

A stepwise procedure was used for purifying the ability indicator with the logit model and RFA.

Instead of redefining the total test score by taking into account only those items revealed as unbiased by the technique, the ability measure was redefined at each step as the score on all items in the test but the item with the highest statistically significant value of the dif statistic in the previous step; dif indices were then recalculated for the remaining items in a new run, introducing this refined measure as the ability indicator. This process had to be repeated item by item until all items left were detected as unbiased. The program Iterative Item Bias Detection Version 1.0 (Lucassen, 1991) was used for purifying the ability indicator with the logit model.

Results

Table 2 summarizes the results obtained per iteration with the six dif detection techniques considered in the study. It contains the correct identification rate (HIT: number of biased and unbiased items correctly identified divided by total number of items in the test), Type I error rate or false positive classification errors (FP: items that exhibit dif when they truly do not divide by the number of unbiased items in the test) and Type II error rate or false negative classification errors (FN: items that do not exhibit dif when they truly function differentially across groups divided by the number of biased items in the test). Values showed in the table are the mean of the values obtained in the four datasets considered in the study. The first iteration results show how the dif detection technique works without purification, before introducing any purifying mechanism to the ability scale. Most of these techniques purify the ability scale by removing simultaneously all items detected as biased in the first iteration. Second iteration results are displayed in Table 2 for these methods – Mantel Haenszel, logistic regression and IRT-based techniques. However, logit model and restricted factor analysis carry out a stepwise purification procedure, removing only one item in each step until all remaining items are unbiased ones. Thus, last iteration results are displayed for these two methods in the table.

As the table clearly shows, the results obtained are in agreement with those found previously in the literature: The iterative procedure introduced to purify the ability scale improved dif detection over the noniterative approach in all dif detection techniques. In all cases, the HIT rate increased greatly, while that of FP and FN decreased and in some instances even disappeared.

These analyses reveal that the best techniques for detecting dif were Mantel Haenszel and logit model procedures, which presented a 100% success rate in all four datasets. However, these were not the techniques that showed the greatest improvement when introducing a purifying mechanism, but rather IRT-based indices. First iteration HIT rates for both SA and SOS1 indices were relatively low (between 62% and 70%), but, second iteration rates increased considerably to 97–99%. Even more spectacular is the decrease in FP rates in both indices when introducing the purifying mechanism: from 48–57% to 2–4%. This trend is also shown by the FN rate, though the decrease and the initial rates are considerably smaller. Logit model is also a technique in which

the introduction of a purifying mechanism has a great impact on dif detection: HIT rate rose from 73% to 100%, so classification errors disappeared completely. FP rate was quite high in the first iteration (43%).

The worst results were obtained with logistic regression procedure. In this case, the HIT rate was 93%, the FP rate was quite low (2%), but the FN rate was still 15% (half the initial rate). In spite of this, second iteration results revealed an important improvement relative to those obtained in the first iteration. Having obtained worse results with this technique than with the others might be related to the type of dif simulated in the present study, that is, uniform dif. When only uniform dif is present, one degree of freedom is lost unnecessarily, and this may adversely affect the power of the procedure (Swaminathan & Rogers, 1990).

The purifying mechanism also has a positive effect on restricted factor analysis. The HIT rate increased from 71% to 96%, the FN rate disappeared completely, and the FP rate fell from 36% to 7%.

Table 3 contains the number of items in the refined measure, as used in the last iteration for each technique and dataset. This table clearly portrays the high rate of classification errors made in the first iteration with IRT-based procedures. In fact, the refined criterion should have 15 unbiased items, and in this case, it only has 5–11 items because of the high number of FP errors; further, some of them (1–2) are biased. The refined ability scale defined with the remaining techniques is composed of 11–15 items, some of which are also biased items (particularly in the case of logistic regression), but the final number of items included in the refined measure is closer to the target number. If we now take into account the information conveyed by Tables 2 and 3 together, we can conclude that the techniques that produce the best refined criterion are the stepwise procedures, that is, logit model and restricted factor analysis. In these cases, the number of items included in the purified ability measure is very close to 15 and, what is more important, almost all of them are truly unbiased items. Thus, these stepwise procedures achieve an unbiased measure of ability that can be safely used to investigate the differential functioning of the items of the test.

These results raise two interesting points. First, despite the considerably shortened measure of ability, IRT-based procedures achieve a degree of precision in dif identification similar to that obtained with other techniques with longer refined measures (about 100% HIT rate). Second, the great decrease in the number of items eventually included in the refined measure might have a potential impact on its validity as a measure of the target ability. Do the initial and final scales measure the same ability? Undoubtedly, this issue deserves further consideration.

Table 3. Number of items in the refined criterion.

	Dataset			
	1	2	3	4
Mantel Haenszel	11	12	14	14
Logit model	15	15	15	15
Logistic regression	15	14	15	15
Restricted factor analysis	14	14	14	13
Signed $area_{significance\ test}$	8	8	11	8
Signed $area_{baseline}$	8	8	9	7
Sum of squares 5	7	10	8	

The main results herein obtained may be summarized as follows:

1) When "purifying' the ability measure used, differences in dif identification were seldom found among the techniques considered, each presenting a HIT rate very close to 100%.
2) The best results were obtained with Mantel Haenszel, logit model, and IRT-based techniques, while the worst – although they could still be qualified as good – with logistic regression procedure.
3) The greatest impact of ability purification was observed on IRT-based indices, which were the techniques that improve most as a result of the introduction of this mechanism.
4) Stepwise procedures seemed to produce better refined ability measures than the other two-stage techniques.
5) FP classification errors were more frequent than those of FN in all techniques considered, with the exception of logistic regression.
6) The techniques that showed the highest rate of classification errors in the first iteration were IRT-based procedures.

Discussion

In the light of these results we conclude that ability scale purification is an advisable practice when investigating dif, regardless of the technique used, at least when working with conditions similar to those simulated in this study. Particularly, the ability measure should be purified whenever IRT-based techniques are to be used. This must be the case when we want to guarantee an adequate and true rate of correct identifications. If, for whatever reason, such a purifying mechanism could not be introduced, then these techniques should not be used, and the procedure chosen should be Mantel Haenszel, since this shows the highest HIT rate and the lowest FP and FN rates in the first iteration, without purification. Although Mantel Haenszel, logit model, and IRT-based indices

provide similar results when purifying the ability measure, overall results herein obtained suggest that Mantel Haenszel has the greatest advantages as a dif detection technique, since it is the simplest and it provides the best results without purification.

An interesting result also found in this investigation is that stepwise procedures such as those used with the logit model and restricted factor analysis produced better refined ability scales than procedures that removed in a single step all items revealed as biased by the previous analysis, that is, Mantel Haenszel, logistic regression and IRT-based procedures. This might mean that if the removal of biased items were carried out step by step (one item per step) in the latter techniques, a better refined measure might be obtained. That is, it would be possible to investigate dif with truly unbiased measures. Gómez and Navas (1996) investigated this possibility when studying dif with logistic regression. The results obtained encourage the stepwise removal of biased items, although the computational burden is heavy.

This is just one of the lines of research suggested by the present study. At least, two more could be proposed, both related to the particular conditions of the study.

First, the study works with simulated data. However, issues of validity cannot be addressed in a simulation study. As we have already pointed out, validity may be threatened when purifying the ability measure, since this is modified when biased items are removed. We need to consider carefully whether refined criterion really measures the same ability as the original criterion, and this must be done with real data. The question to be addressed is the following: We are finally able to match examinees using an unbiased measure, but of what?

Second, the present study offers some insights into the nature of the effect of introducing various purifying procedures on the ability measure when studying dif in a very specific situation: simulated data with equal sample size ($N = 1000$) and equal θ distribution for the reference and focal groups, with only uniform dif, a high percent of biased items (40%) and large dif ($b_F - b_R = 0.75$) in a test of only moderate length ($n = 25$). Much more research is needed, however, if we want to generalize these results. Additional research under conditions different from those simulated in the study should be carried out in order to generalize the results obtained. Special attention should be paid to factors such as percent of biased items and test length, since these factors have an important impact on the significance of classification errors and, thus, on the appropriateness of refining the ability measure. Undoubtedly, the present study has shown that ability scale purification has a positive effect on identification of dif with large samples, large dif and short tests. This effect is apparent in all the techniques considered here, to a greater or lesser extent. However, this effect could be less evident if factors such as the two just mentioned – percent of biased items and test length – were modified. It would be worthwhile to further research these issues because of the computational and analytical burden that a purifying mechanism introduces into the analysis.

Acknowledgments

This study was supported in part by funds from the Direccion General de Enseñanza Superior of the Ministry of Education (PB98-0009).

References

Assessment Systems Corporation (1988). *User's manual for the MicroCAT testing system*. St. Paul, MN: Assessment Systems Corporation.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Beverly Hills, CA: Sage.

Candell, G.L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.

Candell, G.L., & Hulin, C.L. (1986). Cross-language and cross-cultural comparisons: Independent sources of information about item non-equivalence. *Journal of Cross-Cultural Psychology, 17*, 417–440.

Clauser, B., Mazor, K., & Hambleton, R.K. (1993). The effects of purification of the matching criterion on the identification of dif using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269–279.

Cole, N.S., & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational measurement*. New York: Macmillan.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19–29.

Flier, H. van der, Mellenbergh, G.J., Ader, H.J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131–145.

Gómez, J., & Navas, M.J. (1996). Detección del funcionamiento diferencial de los items mediante regresión logística: Purificación paso a paso de la habilidad [Dif detection through logistic regression: Stepwise ability purification]. *Psicológica, 17*, 397–411.

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.

Holland, P.W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hulin, C.L., & Mayer, L. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology, 71*, 83–94.

Jöreskog, K.G., & Sörborm, D. (1993a). *LISREL 8 User's reference guide*. Chicago, IL: SSI.

Jöreskog, K.G., & Sörborm, D. (1993b). *PRELIS 2 User's reference guide*. Chicago, IL: SSI.

Kim, S., & Cohen, A.S. (1992). Effects of linking methods on

detection of dif. *Journal of Educational Measurement, 29*(1), 51–66.

Kok, F.G., Mellenbergh, G.H., & Flier, H. van der (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*, 295–303.

Lautenschlager, G., Flaherty, V.L., & Park, D. (1994). IRT differential item functioning: An examination of ability scale purification. *Educational and Psychological Measurement, 54*(1), 21–31.

Lautenschlager, G., & Park, D. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12*, 365–376.

Lim, R.G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164–174.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: LEA.

Lucassen, W. (1991). *Iterative item bias detection*. Leiden.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.

Miller, M.D., & Oshima, T.C. (1992). Effect of sample size, number of biased items and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*, 381–388.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.

Oort, F.J. (1992). Using restricted factor analysis to detect item bias. *Methodika, VI*, 150–166.

Oort, F.J. (1993). Theory of violators: Assessing unidimensionality of psychological measures. In R. Steyer, K.F. Wender, & K.F. Widaman (Eds.), *Psychometric methodology*. Stuttgart: Gustav Fischer Verlag.

Park, D.G. (1988). *Investigations of item response theory item bias detection*. Unpublished doctoral dissertation, Department of Psychology, University of Georgia, Athens.

Park, D.G., & Lautenschlager, G.J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14*, 163–173.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

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

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105–116.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213–233.

Scheuneman, J.D., & Bleistein, C.A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2*, 255–275.

Segall, D.O. (1983). *Test characteristic curves, item bias and transformation to a common metric in IRT: A methodological artifact with serious consequences and a simple solution*. Unpublished manuscript, University of Illinois, Department of Psychology.

Shepard, L.A., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93–128.

Shepard, L.A., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*(2), 77–105.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

María J. Navas-Ara
UNED
Facultad de Psicología
Departamento de Metodología de las Ciencias del Comportamiento
Ciudad Universitaria, s/n
E-28040 Madrid
Spain
Tel. +34 1 398-6235
Fax +34 1 398-7748
E-mail mjnavas@psi.uned.es