DOCUMENT RESUME

ED 476 924 TM 034 975

AUTHOR	Bertrand, Richard; Boiteau, Nancy
TITLE	Comparing the Stability of IRT-Based and non IRT-Based DIF Methods in Different Cultural Contexts Using TIMSS Data.
PUB DATE	2003-00-00
NOTE	20p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Cross Cultural Studies; *Cultural Differences; Foreign Countries; Item Bias; Item Response Theory; Mathematics Tests; Reliability
IDENTIFIERS	*Item Bias Detection; Japan; Mantel Haenszel Procedure; *Score Stability; Third International Mathematics and Science Study; United States

ABSTRACT

This study aimed at finding criteria like within-method stability rates or between-method agreement rates that could help to choose a powerful and low-cost differential item function (DIF) detection method. The study tried to verify the within-method stability of item response theory (IRT) based over non-IRT-based procedures in two different cultural contexts. Data from the Third International Mathematics and Science Study for 1995 and 1999 were used to see if the items identified as having translation DIF in 1995 between U.S. and Japanese groups were the same in 1999. Four procedures were used for that purpose: two IRT-based and two non-IRT-based. In each case, absolute and relative criteria were used to classify the strength of the DIF. The study also investigated between-method agreement rates. Results show that the non-IRT-based methods, especially Mantel Haenszel (a low cost method), possessed between-method agreement rates as high as those obtained by IRT-based methods. Also, the stability rates of non-IRT-based methods were found to be very close to the stability rates of IRT-based methods. (Contains 11 tables and 39 references.) (SLD)

Comparing the Stability of IRT-Based and non IRT-Based DIF Methods in Different Cultural Contexts Using TIMSS Data

Richard Bertrand
Nancy Boiteau

Comparing the stability of IRT-based and non IRT-based DIF methods in different cultural contexts using TIMSS data

Richard Bertrand, Université Laval
Nancy Boiteau, Université Laval

Summary

The undertaking of a DIF operation can be costly and time consuming, especially if we were to use two or more DIF detection methods just to be sure we identified the «right» DIF items. This paper aims at finding criteria like within-method stability rates or between-method agreement rates that could help to choose a powerful and low cost DIF detection method.

Boiteau, Bertrand, Frenette & Saint-Onge (2002) showed that in similar cultural contexts (French Canadians, English Canadians), IRT-based DIF procedures were somewhat more stable than non-IRT based procedures from one linguistic group (English) to another (French). In the present study we tried to verify this within-method stability of IRT-based over non IRT-based procedures in two different cultural contexts. TIMSS95 and TIMSS99 data were used to see if the items identified as having translation DIF in 1995 between USA (reference group) and Japan (focal group) were the same in 1999. Four procedures were used for that purpose: two IRT-based procedures, the UPD index (Shepard, Camilli & Williams 1984; Camilli & Shepard, 1994) and the NCDIF index proposed by Raju, van der Linden & Fleer (1995); and two non-IRT based procedures, the Mantel-Haenszel (MH) approach (Holland & Thayer, 1988) and logistic regression (LR) method (Clauser & Mazor, 1998). In each case, absolute and relative criteria were used to classify the strength of DIF. The absolute criteria are those proposed by Ziecky (1993) for MH, by Gierl, Rogers and Klinger (1999) for LR, by Boiteau, Bertrand, Frenette & Saint-Onge (2002) for the UPD index and by Raju, van der Linden & Fleer (1995) for the NCDIF index. The relative criteria are based on outliers detection used in box-and-whiskers diagram (Tukey, 1977). This paper also investigated between-method agreement rates. Results show that non IRT-based methods and especially Mantel-Haenszel (a low cost method) possessed between-method agreement rates as high as those obtained by IRT-based methods. Also, the stability rates of non-IRT based methods have been found to be very close to the stability rates of IRT-based methods: this last result challenged the one obtained by Boiteau, Bertrand, Frenette & Saint-Onge (2002) and Boiteau & Bertrand (in press) since they found IRT-based procedures somewhat more stable.

Test translation/adaptation issue

The globalization context which prevails today also hit the assessment arena (O'Leary, 2002). More and more tests must now be translated/adapted from a language/culture to another. Many have raised the issue of lack of measurement equivalence for tests translated/adapted from a language/culture to another (Allalouf, 2003; Hambleton, 1993; Poortinga, 1995; Sireci, 1997). The International Test Commission developed guidelines to take account of that very important issue (Arnold et Matus, 2000 ; Hambleton, 2001). International large-scale assessments like the Third International Mathematics and Science Study (TIMSS) or the Program for International Student Assessment (PISA) involving tests translated in many languages must take this problem very seriously.

Problems of different types are associated with this translation/adaptation process in large-scale settings. Among these numerous problems is the credibility of comparisons involving countries that differ both linguistically and culturally (Hambleton, 1993; O'Leary, 2002 ; Sireci, 1996 ; Wainer, 1994). To give these comparisons more credibility a very rigourous translation process must be followed. But even in this situation, it would be rash to suppose that the source and the target instruments are necessarely equivalent (Ercikan, 1999; Poortinga, 1995;

Van de Vijver & Hambleton, 1996). In fact, most of the time this process produce what can be called **translation bias**. As Hambleton (2001) noted, use of judgmental reviews is not enough; empirical studies must be undertaken to identify and control for those translation bias.

In the last decades, quite a few statistical procedures have been proposed to identify translation bias. Van de vijver & Leung (1997) developed a three-way classification to identify potential sources of bias: construct bias, method bias and item bias. Multiple statistical procedures were proposed to detect construct bias (Van de vijver & Leung, 1997; Gierl, 2000; Bertrand *et al.*, 2001) and method bias (Hambleton, 2001; Bertrand *et al.*, 2001) but most of the statistical procedures were developed to control item bias, the so-called differential item functioning (DIF) procedures. Now these procedures don't always agree perfectly well: some are more liberal, some are more conservative; some are «cheap», some are expensive. How are we going to choose between these procedures? The present authors think that between-method agreement rates and within-method stability rates should be considered for that purpose.

The main purpose of this paper is to compare between-method agreement rates and within-method stability rates of IRT-based methods and non IRT-based methods to detect DIF items using the 1995 and the 1999 TIMSS math assessments. Boiteau, Bertrand, Frenette & Saint-Onge (2002) showed that in similar cultural contexts (French Canadians, English Canadians), IRT-based DIF procedures were somewhat more stable than non-IRT based procedures. Besides, they found fair between-method agreement rates. But, as noted by Candell & Hulin (1986), Hambleton & Kanjee (1995) and Hambleton (2001), the cultural distance must be taken into account in DIF studies. Therefore, this paper aims at verifying the between-method agreement rates and the within-method stability rates of IRT-based over non IRT-based procedures in two very different cultural contexts (Japan and USA) using the 1995 and 1999 TIMSS math data sets.

Procedures

Only the 48 TIMSS math items common to the 1995 and the 1999 assessments were used in this study. These 48 items can be found into one or the other of three of the TIMSS booklets (1, 5 and 7). We ended up, for each booklet and each group, with samples of about 1300 students for the 1995 assessment; for the 1999 assessment we had, for each booklet, about 1100 students for the reference group and 600 students for the focal group.

The dimensionality of the scales were tested using full information item factor analysis (Bock, Gibbons & Muraki, 1988) as implemented in TESTFACT4 (Bock, Gibbons, Shilling, Muraki, Wilson & Wood 2003). A TESTFACT analysis was performed for each the three booklets and each of the linguistic groups.

Next, we decided not to use a multistage procedure to purify[1] the internal criterion (ability) partly because this procedure was too costly and time consuming. Besides, while some authors (Navas-Ara & Gomez-Benito, 2002; Zenisky, Hambleton & Robin, 2003) would argue in favor of this procedure, some (Gierl, Jodoin & Ackerman, 2000) would are not convinced.

Since the statistical tests used in the DIF detection methods are affected by sample size, we decided to use a **relative criterion**, besides the **absolute criterion** described below, to identify and classify DIF items. To this end, the box-and-whiskers plot (see Figure 1) was used to examine the outlier and extreme values of the statistic involved (UPD, NCDIF, $|\Delta|$, χ^2). An extreme value on the plot (located at more than 3 times the width of the interquartile range from the 3rd quartile) indicated a **severe** DIF (category C). An outlier (not extreme) value (located at more than 1.5 times but less than 3 times the width of the interquartile range from the 3rd quartile)

[1] Preliminary results using two-stage purifying procedure for non IRT-based methods were very consistent with the results obtained with the procedure chosen here (not using purified internal criterion).

indicated the presence of a **moderate** DIF (category B). Otherwise a trivial or negligible DIF was supposed (category A).

Insert Figure 1 about here

Non based IRT DIF methods

Mantel-Haenszel

Holland and Thayer (1986) proposed a statistic, previously discussed by Mantel and Haenszel (1959), to develop a method for detecting DIF. Throughout the years, this method became more and more popular (Zwick, 1997). The Mantel-Haenszel method compares, for a given item, the probability of obtaining a right answer in the focal group to the probability of obtaining a right answer in the reference group for subjects of equal ability.

There are many ways to determine the presence of DIF using the α_{MH} statistic. The one used in the present paper became a favorite to common users of the Mantel-Haenszel method (Roussos et al., 1999). It allows for a more complete interpretation of DIF items. First, the value $\Delta_{MH} = -2.35 \ln(\alpha_{MH})$ is obtained. Negative Δ_{MH} values correspond to items favoring the reference group. According to Ziecky (1993) if the absolute value of Δ_{MH} is higher than 1.5 and significantly higher than 1 (at $\alpha = .05$), the item is classified as category C (severe DIF). If the absolute value of Δ_{MH} is lower than 1 or not significantly higher than 0 (at $\alpha = .05$), the item is classified as category A (trivial DIF). In all other situations, the item is classified as category B (moderate DIF).

Logistic regression

Logistic regression (Swaminathan & Rogers, 1990) allowed for the development of a now very popular DIF detection method (Clauser & Mazor, 1998). The logistic regression procedure involves two stages. In the first stage, total test score is included in the regression equation. In the second stage, two variables related to the group and the interaction group*score, are included in the equation. The analysis consists in testing if the inclusion of these two variables leads to a statistically significant verdict. If so, it can be said that the item is DIF.

The absolute criterion used here is the one proposed by Gelin & Zumbo (2003) and Jodoin & Gierl (2001). An item would be considered to possess a severe DIF if the chi-square test associated with the second stage is found statistically significant and if the R-square difference between the two stages is higher than 0.07. An item would be considered to have a moderate DIF if the chi-square test associated with the second stage is found statistically significant and if the R-square difference between the two stages is higher than 0.035 but less than 0.07. In all other cases, DIF is considered trivial.

IRT- based DIF methods

The area method (UPD index)

The area method (Shepard, Camilli et Williams, 1984; Camilli & Shepard, 1994) focus on a quantity that reflects the difference between the reference group and the focal group ICC's. Two indices were proposed to that end: a

signed index (SPD-θ^2) and an unsigned (UPD-θ^3) index. If the two ICC's cross, the difference of probabilities involved in the computation of the signed index can cancel out and the value of this index can be low even if a large or moderate DIF is manifestly present. Since in this study we want to detect uniform as well as non-uniform DIF we will rely on the UPD index which values are always positive. The signed index would be useful if we were interested in identifying which group was favored by the item. Notice that the sum of the difference of probabilities is taken over the number of subjects in the focal group (n_F).

$$\text{SPD-}\theta = \Sigma_j[\, P_{iR}(\theta_j) - P_{iF}(\theta_j)] \,/\, n_F \qquad \text{where } j = 1, 2, \ldots, n_F.$$

$$\text{UPD-}\theta = \left(\Sigma_j[\, P_{iR}(\theta_j) - P_{iF}(\theta_j)]^2 \,/\, n_F\right)^{.5} \quad \text{where } j = 1, 2, \ldots, n_F.$$

Since we don't know any absolute criterion related to the UPD index, we decided to use the value of .10 as a threshold: this amounts to consider DIF items those for which the overall difference of probabilities between the focal group and the reference group is higher than .10.

Raju's NCDIF index

Raju, van der Linden & Fleer (1995) proposed a very refined framework to look at DIF items. Following these authors, two approaches are possible. The first one looks at items that can prevent valid comparisons using total score to compare the focal group and the reference group. The second one is interested in identifying DIF items that could offense subgroups (Blacks, girls, handicapped, etc.) of a population and that must be changed or else completely removed from the item bank.

The first approach involves a differential test functioning index (DTF) and a compensatory or signed DIF index (CDIF) for each item. Since the sum of the values of the CDIF for all items in the test is equal to the DTF value, the procedure implies the identification and removal of items (one at a time) with the largest and positive CDIF values until the DTF index is no more statistically significant.

It can be shown that

$$DTF = \varepsilon_j(D_j^2) = \sigma^2_{D_j} + \overline{D_j^2}$$

$$\text{DTF} = \Sigma_i \text{ CDIF}_i$$

$$CDIF_i = \varepsilon_j(d_{ij} D_j) = \sigma_{d_{ij} D_j} + \overline{d_{ij}}\ \overline{D_j}$$

where $d_{ij} = P_{iR}(\theta_j) - P_{iF}(\theta_j)$ and $D_j = V_R(\theta_j) - V_F(\theta_j)$ for item i and ability level θ_j.

The second approach involves a non-compensatory (unsigned) DIF index (NCDIF) used in the same sense as the UPD index described above.

The NCDIF index is given by the following formula:

[2] Following Camilli & Shepard (1994, p.67) this reads signed probability difference controlling for theta.
[3] Unsigned probability difference controlling for thêta.

$$NCDIF_i = \varepsilon_j(d_{ij}^2) = \sigma_{d_{ij}}^2 + \overline{d_{ij}^2}$$ where $d_{ij} = [P_{iR}(\theta_j) - P_{iF}(\theta_j)]$ and $j = 1, 2, ..., n_F$.

and n_F refers to the number of subjects in the focal group.

Chi-square test associated with (n_F degrees of freedom) this statistic is given by

$$\chi^2 = \frac{n_F * NCDIF_i}{\sigma_{d_{ij}}^2}$$

The NCDIF index is non compensatory and non signed which means that the values of this index are always positive. This index tends to identify items for which the area between the two ICCs is large. In accordance with this method (McCarty, Oshima & Raju, 2002; Raju, van der Linden & Fleer, 1995), and since the chi-square statistic is influenced by sample size, an item is judged as presenting DIF if the value of NCDIF is higher than 0.006 **and** if the chi-square value leads to a statistically significant verdict (at $\alpha = .01$).

Results in a Canadian (common culture) context

Using the data from the Canadian School Achievement Indicators Program (SAIP), Boiteau, Bertrand, Frenette & Saint-Onge (2002) showed that in similar cultural contexts (French Canadians, English Canadians), IRT-based DIF procedures were somewhat more stable than non-IRT based procedures from one linguistic group (English) to another (French). They used samples of 20 000 students in each of the 1996 and 1999 SAIP science assessments.

Table 1 shows that IRT-based methods (NCDIF, UPD) were found more stable than non IRT-based methods (MH, LR). Using the 29 items identified as DIF by at least one method, the stability rates of the IRT-methods were found higher than 75%, that is more than 75% of the decisions (this item is considered DIF or not!) taken in the 1996 SAIP assessment were the same in the 1999 SAIP assessment.

Insert Table 1 about here

Results obtained by Boiteau, Bertrand, Frenette & Saint-Onge (2002) have also shown that MH seemed to produce lower between-method agreement rates than LR, NCDIF or UPD.

In another study involving more than 20 000 students in each of the 1997 and the 2001 SAIP math assessments, Boiteau & Bertrand (in press) concluded (table 2) that non IRT-based methods were somewhat less stable than the IRT-based methods.

Insert Table 2 about here

Results in foreign (different culture) contexts

Within-method stability rates

Table 3 summarized the results obtained from the 1995 and the 1999 TIMSS assessments (booklets 1, 5, 7). It can be seen that the overall within-method stability rates involving non IRT-based methods are very much the same as the rates associated with IRT-based methods. These rates are quite high as they are all close to 80%.

Insert Table 3 about here

A look at table 4 (booklet 1), table 5 (booklet 5) and table 6 (booklet 7) shows that large DIF (item 3, item 8, item 11) are detected by almost all methods in the three booklets.

Insert Table 4, 5 and 6 about here

Between-method agreement rates

As seen in table 7, the two IRT-based methods (NCDIF, UPD) got a very high agreement rate (92%) while the two non IRT-based methods (MH, LR) got a quite low agreement rate (67%). Overall, the two methods that had the highest agreement rate (95%) are MH and NCDIF.

Insert Table 7 about here

Analyzing each booklet separately for each assessment (tables 8 through 13) it can be seen that between-method agreement rates were generally high especially for the IRT-based methods. While MH produced agreement rates as high as NCDIF and UPD, the between-method agreement rates were not so good for LR, and especially those related to booklets 5 and 7.

Insert Tables 8 to 13 about here

Discussion

The aim of this study was to compare within-method stability rates and between-method agreement rates of four DIF detection methods: two IRT-based methods, NCDIF and UPD, and two (low cost) non IRT-based methods, MH and LR. Data from USA and Japan samples from the 1995 and 1999 TIMSS math assessments were used for that purpose.

While we found in other studies (Boiteau, Bertrand, Frenette & Saint-Onge, 2002; Boiteau & Bertrand, in press) that IRT-based methods were somewhat more stable, this study shows that non IRT-based methods are as stable as IRT-based. Many reasons can explain the lack of convergence of these results. First, the former studies were undertaken in similar cultural contexts while the present study compares two very different cultures/country, Japan and USA. Second, the former studies involved a lot less large DIF items: only about 20% of the items were then detected as DIF. On the other hand, in our present study, most of the 48 items that we worked with were classified as moderate or severe DIF by at least one method. Also, we did not use internal criterion (total score) purification (Navas-Ara & Gomez-Benito, 2002) that may have generated different results, especially for IRT-based methods. Third, the fact that we got many large DIF in the present study makes it easier for all methods to detect those DIF items and therefore to have high within-method stability rates as well as high between-method agreement rates.

The lack of purification of the internal criterion could be considered a limitation of this study but we should keep in mind that

- results from table 2 and table 3 based on two different studies are consistent;
- a preliminary study using purified internal criterion (two-stages approach) for non IRT-based methods showed consistent results with what we got here not using purified internal criterion;
- Gierl, Jodoin & Ackerman (2000) argued that purification may be unnecessary for methods like LR;
- Zinesky, Hambleton & Robin (2003) specified that although the purification of the internal criterion is a well known issue, most of the researchers don't use it anyway;
- Navas-Ara & Gomez-Benito (2002) arguing for a purified internal criterion mentioned that, even without purifying, results from MH were relevant: our results show that MH got very high agreement rates with other methods, and especially with IRT-based methods.

Based on the present results, we can argue that MH got much higher between-method agreement rates than LR. UPD method also performed very well. Now this method has a quite interesting intuitive appeal: an item is said to be DIF if the overall probability difference between the ICC of the focal group and the ICC of the reference group is higher than .10. Overall, methods used here got much better between-method agreement rates than observed in similar situations. Price (1999) for example found an agreement rate of only[4] 20% between NCDIF and MH using also tests translated from English to Japanese. Gierl, Rogers & Klinger (1999) though found an agreement rate of 90% between LR and MH using a Canadian math test (reference group was English and focal group was French Immersion) while we found only a 67% agreement rate between those two methods. None of these studies used purification of the internal criterion.

Some of our results are also consistent with other studies. For example, using logistic regression as a DIF detection procedure, Ercikan (1999) found a little more than 18% of DIF items in math TIMSS items (1995); we found, also using logistic regression, 9/24 (38%) DIF items in booklet 1, 5/24 (21%) DIF items in booklet 5 and 4/24 (16%) DIF items in booklet 7.

[4] Price reported 20% agreement rate but our analysis of his data showed 73% agreement rate using our definition.

A final result is worth reporting: we found that all but one item flagged as an outlier/extreme value detected by a box-and-whisker (relative criterion) plot were also detected by the Raju's DTF-CDIF procedure. Our study showed also that this relative criterion tends to detect much less DIF items when the DIF statistic involved (UPD, NCDIF, $|\Delta|$, χ^2) had large variance, that is when the box width was large.

(Many) more studies are surely needed before selecting «the best» (powerful, low cost, low type I error rate, high within-method stability rates, high between-method agreement rates) detection method, whether IRT-based or non IRT-based. Among those, the usefulness of a purified internal criterion for the IRT-based methods should be investigated thoroughly.

References

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in education, 16, 1*, 55-73.

Arnold, B.R., & Matus, Y.E. (2000). Test translation and cultural equivalence methodologies for use with diverse populations. In I. Cuellar & F. A. Paniagua (eds.), *Handbook of multicultural mental health, 2* (pp. 121-136). San Diego, Ca: Academic Press.

Bertrand, R., Boiteau, N., Gauthier, N., Compain, C., Frenette, É., Laprise, A., Léger-Bourgoin, N., & Jeanrie, C. (2001). *La gestion des biais de concept, des biais de méthode et des biais d'item dans le contexte des enquêtes du Programme des indicateurs de rendement scolaire (PIRS)*. Document interne. Québec : Université Laval.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R. D., Gibbons, R., Shilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT4. Test scoring, item statistics and item factor analysis. Mooresville, IN : Scientific Software.

Boiteau, N., Bertrand, R., Frenette, É., & Saint-Onge, C. (2002). *Stability of IRT-based and non IRT-based DIF procedures*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Boiteau, N., & Bertrand, R. (in press). Apport de la théorie de la réponse à l'item dans la stabilité des méthodes de détection d'un fonctionnement différencié d'item. In Blais, J-G. et Raîche, G. (eds). *Regards sur la modélisation de la mesure en éducation et en sciences sociales*. Québec: Presses de l'Université Laval

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

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

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement : Issues and practices*, Spring , 31-44.

Ercikan, K. (1999). *Translation DIF in TIMSS*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Montréal.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the Center for epidemiologic studies depression scale. *Educational and Psychological measurement, 63, 1*, 65-74.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*, 4, 280-296.

Gierl, M. J., Rogers, W.T., & Klinger, D.A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *The Alberta Journal of Educational Research, 45*, 4, 353-376.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, Simultaneous item bias test and logistic regression when the proportion fo DIF items is large*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of psychological assessment, 9,* 57-68.

Hambleton, R.K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: use of improved methods for test adaptations. *European Journal of psychological assessment*, 11, 3, 147-157.

Hambleton, R.K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, **17, 3**, 164-172.

Holland, P. W., & Thayer, D.T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. Technical Report. Princeton, NJ: Educational Testing Service.

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

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in education, 14*, 4, 329-349.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Mc Carty, F. A., Oshima, T. C., & Raju, N. (2002). Identifying possible sources of differential functioning using differential bundle functioning with polytomous scored data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, **18, 1**, 9-15.

O'Leary, M. (2002). Stability of country rankings across item formats in the TIMSS. *Educational Measurement: Issues and Practice, 21*, 27-38.

Poortinga, Y. H. (1995). Cultural bias in assessment: historical and thematic issues. *European Journal of psychological assessment, 11*, 3, 140-146.

Price, L. P. (1999). *Differential functioning of items and tests versus the Mantel-Haesnzel technique for detecting differential item functioning of a translated test.* Paper presented at the annual meeting of the American Alliance of Health, Physical Education, Recreation and Dance, Boston.

Raju, N.S, van der Linden, W.J., & Fleer, P.F. (1995). IRT-based internal measure of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 4, 353-368.

Roussos, L.A., Schnipke, D.L., & Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functionning parameter. *Journal of Educational and Behavioral Statistics, 24, 3,* 292-322.

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

Sireci, S. G. (1996). *Technical issues in linking assessments across languages.* Paper presented at the annual meeting of the National Council on Measurement and Evaluation (NCME), New York.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*, 12-19.

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

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, Mass: Addison-Wesley.

Van De Vijver, F., & Hambleton, R.K. (1996). Translating tests: some practical guidelines. *European psychologist*, **1, 2**, 89-99.

van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Wainer, H. (1994). Problèmes de mesure. *Mesure et évaluation en éducation, 17,* 2, 115-146.

Zenisky, A. L., Hambleton, R. K., & Robin. F. (2003). Detection of differential item functioning in large-scale state assessments: a study evaluating a two-stage approach. *Educational and ,Psychological Measurement, 63*, 1, 51-64.

Ziecky, M. (1993). Practical questions in the use of DIF statistics in item development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and Psychological Measurement*, 57, 3, 412-421.

Figures

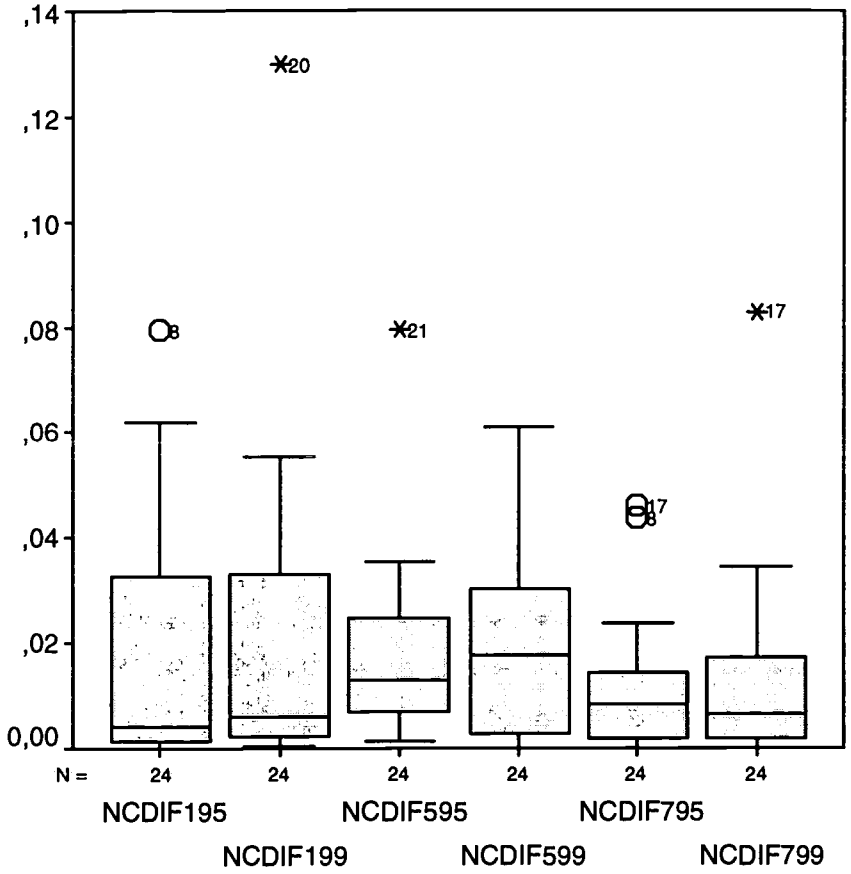

Figure 1 Box-and-whiskers of NCDIF index for booklets 1, 5 and 7, using the 95 and 99 TIMSS assessements showing outlier (O) and extreme (*) values

Tables

Table 1 Within-method stability rates of DIF methods using the 1996 and 1999 SAIP science assessments : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	Non IRT-based methods				IRT-based methods			
Items	MH		LR		NCDIF		UPD	
	96	*99*	*96*	*99*	*96*	*99*	*96*	*99*
# stable DIF	15/29		19/29		22/29		26/29	
Stability rate	52%		66%		76%		90%	

Table 2 Within-method stability rates of DIF methods using the 1997 and 2001 SAIP math assessments : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	Non IRT-based methods				IRT-based methods			
Items	MH		LR		NCDIF		UPD	
	97	*01*	*97*	*01*	*97*	*01*	*97*	*01*
# stable DIF	10/19		14/19		16/19		17/19	
Stability rate	52%		74%		84%		89%	

Table 3 Overall within-method stability rates of DIF methods using the 1995 and 1999 TIMSS math assessments (booklets 1, 5 and 7): Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	Non IRT-based methods				IRT-based methods			
Items	MH		LR		NCDIF		UPD	
	95	*99*	*95*	*99*	*95*	*99*	*95*	*99*
# stable DIF	58/72		59/72		57/72		61/72	
Stability rate	81%		82%		79%		85%	

Table 4 Within-method stability rates using the 1995 and 1999 TIMSS math assessments (booklet 1) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index. Presented are category C and category B items (X items are also DIF).

	Non IRT-based methods				IRT-based methods			
Items	MH		LR		NCDIF		UPD	
	95	*99*	*95*	*99*	*95*	*99*	*95*	*99*
1	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-
3	C	C	B	B	X	X	X	X
4	-	-	-	-	-	-	-	-
5	-	B	-	-	-	X	-	X
6	-	-	-	-	-	-	-	-
7	B	B	-	-	-	X	-	-
8	C	C	C	C	X	X	X	X
9	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-
11	C	B	B	-	X	X	X	-
12	B	C	-	-	X	X	X	X
13	C	C	B	B	X	X	X	X
14	-	B	-	-	-	-	-	-
15	-	-	-	-	-	-	-	-
16	C	B	B	-	X	-	X	-
17	-	-	-	-	-	-	-	-
18	C	C	C	B	X	X	X	X
25	C^{O}	C	C	B	X	X	X	X
26 (DFT99)	C	C	B	C^{O}	X	X^{O}	X	X
27	-	B	-	-	-	X	-	-
28	-	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-	-
30	C	C	B	-	X	X	X	X
# stable DIF	21/24		21/24		20/24		21/24	
Stability rate	88%		88%		83%		88%	

[X^{O} : This item is also an outlier in a box-and-whiskers plot]

[26 (DTF99) : Item 26 was also detected as DIF in 99 by the DTF-CDIF framework]

Table 5 Within-method stability rates using the 1995 and 1999 TIMSS math assessments (booklet 5) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index Presented are category C and category B items (X items are also DIF).

	Non IRT-based methods				IRT-based methods			
Items	MH		LR		NCDIF		UPD	
	95	*99*	*95*	*99*	*95*	*99*	*95*	*99*
1	C	B	-	-	X	X	X	-
2	-	-	-	-	-	-	-	-
3	C	C	-	-	X	X	X	X
4	B	-	-	-	X	-	X	-
5	B	C	B	B	X	X	X	X
6	-	-	-	-	-	-	-	-
7	-	C	-	B	-	X	-	X
8	C	B	B	-	X	X	X	X
9	-	-	-	-	-	-	-	-
10	B	C	-	-	X	X	X	X
11	C	C	-	-	X	X	X	X
12	-	B	-	-	-	X	-	-
31	-	B	-	B	X	X	X	X
32	C	C	B	-	X	X	X	X
33	-	-	-	-	X	-	-	-
34	B	C	-	B	X	X	X	X
35	-	-	-	-	-	-	-	-
36	C	C	-	B	X	X	X	X
37	B	B	-	-	X	X	X	X
38(DTF95)	C	C	B	B	X	X	X	X
39(DTF95)	C	C	C^{O}	B	X^{O}	X	X^{O}	X
40	C	B	-	-	X	X	-	X
41	C	C	-	B	X	X	X	X
42	B	-	-	-	X	-	-	-
# stable DIF	19/24		17/24		19/24		20/24	
Stability rate	79%		71%		79%		83%	

[X^{O} : This item is also an outlier in a box-and-whiskers plot]

[38 (DTF95) : Item 38 was also detected as DIF in 95 by the DTF-CDIF framework]

Table 6 Within-method stability rates using the 1995 and 1999 TIMSS math assessments (booklet 7) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index

Presented are category C and category B items (X items are also DIF).

	Non IRT-based methods				IRT-based methods			
Items	MH		LR		NCDIF		UPD	
	95	*99*	*95*	*99*	*95*	*99*	*95*	*99*
1	C	-	-	-	X	-	X	-
2	-	-	-	-	-	-	-	-
3	B	C	-	-	X	X	X	X
4	-	-	-	-	-	-	-	-
5	B	C	B	B	X	X	X	X
6	-	-	-	-	-	-	-	-
7	-	B	-	-	-	X	-	-
8(DTF95)	C	C	B^O	B	X^O	X	X	X
9	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-
11(DTF95)	C	C	-	B	X	X	X	X
12	C	B	-	-	X	X	X	X
19	B	C	-	-	X	X	X	X
20	C	C	-	-	X	X	X	X
21	-	-	-	-	-	-	-	-
22	-	-	-	-	-	X	-	-
23(DTF95)	C^O	C^O	C^O	C^O	X^O	X^O	X	X^O
24	B	-	-	-	-	-	-	-
43	-	B	-	-	-	X	-	X
44	C	C	-	-	X	X	X	-
45	C	-	B	-	X	-	-	-
46	B	-	-	-	X	-	X	-
47	-	-	-	-	-	-	-	-
48	C	C	-	C	X	X	X	X
# stable DIF	18/24		21/24		18/24		20/24	
Stability rate	75%		88%		75%		83%	

[X^O : This item is also an outlier in a box-and-whiskers plot]

[23 (DTF95) : Item 23 was also detected as DIF in 95 by the DTF-CDIF framework]

Table 7 Overall between-method agreement rates in the 1995 and the 1999 TIMSS math assessments (booklets 1, 5, 7) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD
MH	67%	95%	88%
LR	-	68%	75%
NCDIF		-	92%

Table 8 Between-method agreement rates in the 1995 TIMSS math assessment (booklet 1) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD	Mean
MH	22/24 92%	23/24 96%	23/24 96%	95%
LR	-	23/24 96%	23/24 96%	95%
NCDIF		-	24/24 100%	97%
UPD				97%

Table 9 Between-method agreement rates in the 1999 TIMSS math assessment (booklet 1) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD	Mean
MH	16/24 67%	22/24 92%	19/24 79%	79%
LR	-	18/24 75%	21/24 88%	77%
NCDIF		-	21/24 88%	85%
UPD				85%

Table 10 Between-method agreement rates in the 1995 TIMSS math assessment (booklet 5): Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD	Mean
MH	13/24 54%	22/24 92%	21/24 88%	78%
LR	-	11/24 46%	14/24 58%	53%
NCDIF	-	-	21/24 88%	75%
UPD				78%

Table 11 Between-method agreement rates in the 1999 TIMSS math assessment (booklet 5) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD	Mean
MH	15/24 63%	24/24 100%	22/24 92%	**85%**
LR	-	15/24 63%	17/24 71%	**66%**
NCDIF	-	-	22/24 92%	**85%**
UPD			-	**85%**

Table 12 Between-method agreement rates in the 1995 TIMSS math assessment (booklet 7) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD	Mean
MH	14/24 58%	23/24 96%	22/24 92%	**82%**
LR	-	15/24 63%	14/24 58%	**60%**
NCDIF		-	23/24 96%	**85%**
UPD				**82%**

Table 13 Between-method agreement rates in the 1999 TIMSS math assessment (booklet 7) : Mantel-Haenszel (MH), logistic regression (LR), NCDIF index and UPD index.

	LR	NCDIF	UPD	Mean
MH	17/24 71%	23/24 96%	22/24 79%	**86%**
LR	-	16/24 67%	19/24 79%	**72%**
NCDIF		-	21/24 88%	**83%**
UPD				**86%**

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034975

I. DOCUMENT IDENTIFICATION:

Title: COMPARING THE STABILITY OF IRT-BASED AND NON IRT-BASED METHODS IN DIFFERENT CULTURAL CONTEXTS USING TIMSS DATA

Author(s): RICHARD BERTRAND & NANCY BOITEAU

Corporate Source: UNIVERSITÉ LAVAL

Publication Date: 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature:	Printed Name/Position/Title: RICHARD BERTRAND, FULL PROFESSOR
Organization/Address: FACULTÉ DES SCIENCES DE L'ÉDUCATION UNIVERSITÉ LAVAL QUÉBEC (CANADA) G1K 7P4	Telephone: 418-656-5089 FAX: 418-656-7770
	E-Mail Address: RICHARD.BERTRAND@FSE.ULAVAL.CA Date: 05/16/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org