COMPARING TRANSLATED SCIENCE ACHIEVEMENT TESTS USING DIFFERENTIAL ITEM FUNCTIONING

By
Ong Saw Lan
osl@usm.my
Lee Lay Wah
School of Educational Studies
Universiti Sains Malaysia

Background Of the Malaysian Education System

Malaysia was a former British colonial country. The Malaysian education system adopted the British Education system which consists of the primary, lower secondary, upper secondary and post secondary. At the primary level, children enter schools at the age of 6 and spend six years from standards 1 to 6. Following this, three years are spent at the lower secondary level (Form 1 to 3) and a further 2 years at upper secondary (Form 4 to 5). Finally two years are spent at the post secondary level (Form Six). All schools follow the same national curriculum. The Ministry of Education conducts national level assessment to measure student achievement in Primary Six, Lower secondary, upper secondary and pre-university level.

There are three major types of primary schools according to the medium of instruction; Malay, Chinese and Tamil. At the end of the six years primary school education, all students will sit for the national level examination. The Examination Board of the Ministry of Education assesses the achievement of these 13-year old students in reading and writing, mathematics, and science using paper-and-pencil tests. Both mathematics and science are being assessed in Malay, Chinese and Tamil according to the medium of instruction. These tests account for a substantial percentage of a student's final grade. The results of these tests are used for making important decisions such as selection for entry into science boarding schools. School administrators include these scores for grouping students into different classes. A lot of publicity is given to the school results as it is reported in local newspapers. Consequently, comparability of tests results across different language version of these tests is an important issue on the validity of interpretation in these assessments.

Issues of tests in different language version

The adaptation and translation of educational tests is becoming more important when increasing number of students are studying in different languages. The international level of testing such as the Third International Mathematics and Science Study (TIMSS) were prepared in 31 languages for the 45 participating countries. In Malaysia, the use of three medium of instruction at the primary school level necessitates the development of tests in Malay, Chinese and Tamil.

The main aim of test adaptation and translation is to ensure maintenance of construct equivalence and content representation across the different language versions (Allalouf, Hambleton & Sireci, 1999). In spite of this, research has shown that translating a test from a source language to a target language does not necessarily produce two psychometrically equivalent tests (Allalouf, Hambleton, & Sireci, 1999; Budgell, Raju & Quartetti, 1995; Ercikan, 1998; Hambleton, 1993). In creating tests that are as similar as possible across different languages, careful translation process is needed to preserves the original meaning of the test. Additional changes in item format may be necessary to ensure equivalence of the test in multiple languages (Hambleton, 1993). The general process of converting one language version of a test to another is known as adaptation (Ercikan et. Al. 2004). Translation is one of the stages in the process of test adaptation across different languages. Poor test adaptation can affect the meaning of test items and relative difficulty. This inadvertently influence the comparability and interpretability of test scores across language groups.

Research on multilingual examinations has demonstrated that test adaptations can affect comparability, and therefore, validity, for groups taking the tests in different languages (e.g., Angoff & Cook, 1988; Sireci & Berberoglu, 2000; Sireci, Fitzgerald, & Xing, 1998; van der Vijver & Tanzer, 1998). Recent research conducted by Ercikan (1998, 1999), Ercikan and McCreith (2002), Gierl, Rogers, and Klinger (1999), and Gierl and Khaliq (2001) using English and French test for Canadian students found psychometric differences between these two language version of tests as well. According to Ercikan et. al. (2004), psychometric differences between the language versions of test due to cultural and curriculum differences between the groups may affect item equivalence across language version of tests. As an example, cultural differences can influence examinees' familiarity with the content or context of items.

Methods Comparing Equivalent Of Different Language Version of Tests

Different methods have been used to compare two different version of the test. The easiest and widely used one is the judgmental method. The involvement of knowledgeable persons in this procedure can help in understanding and identifying causes of DIF, thus providing item writers with guidelines in constructing good items. Some of the studies (Hambleton & Jones, 1995) used judgmental and empirical procedures separately, and checked the level of congruence between them. Researchers found that reviewers are generally poor at predicting which items would function differently across groups (Engelhard, Hansche, & Rutledge, 1990, Gierl & McEwen, 1998; Plake, 1980; Rengel, 1986; Sandoval & Miille, 1980).

When a test is translated from one language into another language, Allalouf (2003) found that the two tests are generally not psychometrically equivalent. Unfortunately, item equivalence across language is often assumed without the use of statistical procedures (van der Vijver & Leung, 1997). Statistical analyses based on differential item functioning (DIF) have been used widely for comparing translated test and adaptation of test between language groups (Gierl and Khaliq, 2001). DIF analysis is a procedure used to identify items that function differently between different groups (Lin & Rogers, 2005). It is based on the underlying assumptions that examinees with similar ability should perform similarly. DIF occurs when an item is substantially more difficult for one group than for another group when the groups ability is taken into consideration (Shepard, Camilli, & Averill, 1981). When used in translated test, DIF technique will detect items function differently across different language groups, if examinees of equal ability but from different language groups do not have an equal probability of responding correctly to that item (Allalouf, Hambleton & Sireci, 1999).

Though statistical analyses are very helpful in detecting DIF items, they do not reveal the causes of the DIF. Hulin (1987) put forth a method based on item response theory (IRT) that could help determine sources of DIF when two different language groups are compared. Hulin suggested the comparison of item characteristic curves of the two language groups be used.

Discrepant item characteristic curves indicate nonequivalence between the two version of tests. Specifically, Hulin (1987) proposed that the item discrimination parameter (a) differences, indicated cultural differences whereas the item difficulty parameter (b) differences indicated translation errors .

Causes of DIF in translated items

Logical analysis carried out on translated items can help to identify the possible sources of DIF. Angoff and Cook (1988) analyzed the equivalence between the Scholastic Achievement Test and its Spanish-language counterpart, the Prueba de Aptitud Academia. They concluded that the amount of text in an item is a significant factor: Items with less text tend to have more translation DIF, whereas items with more text are more likely to retain their meaning (and their psychometric characteristics). Gafni and Canaan-Yehoshafat (1993) and Beller (1995) studied the translation of the Israeli Inter-University Psychometric Entrance Test (PET) from Hebrew into Russian and arrived at the same conclusions as those of Angoff and Cook (1988).

Allalouf, Hambleton, and Sireci (1999) found four main causes for DIF in the translated verbal items. The causes were (a) changes in word or sentence difficulty- the translation resulted some words or sentences became easier or more difficult. (b) Changes in content where the meaning of the item changed in the translation, thus turning it into a different item. (c) Changes in format of the item like a sentence became much longer, or words that originally change in the stem now appeared instead in all four alternative responses., which is due to constraint of the particular language. (d) Differences in cultural relevance where items were exactly the same but the two groups differed because of cultural content of the specific item. This could be due to the content that was more relevant or familiar to one of the groups.

In a study by Gierl and Khaliq (2001), four sources of DIF were identified in Canadian achievement tests administered in English and French. The sources were (a) omission or addition of words or phrases that affect meaning, (b – c) differences in words or expressions inherent and not inherent to the language or culture, and (d) format differences.

AIM OF THE STUDY

The purpose of this paper is to identify item(s) in two versions of science achievement tests that may function differently across language groups. The statistical analyses described by Roussos and Stout (1996) and the difficulty parameter contrast are used to identify and evaluate DIF.

The present paper will address two questions:
(a) How comparable are the two language versions of the science achievement test? and
(b) To what extent the two statistical analyses correspond each other in detecting DIF in science items?

RESEARCH DESIGN

The research design involved two stages. First, detecting DIF using two different statistical analyses. The second stage involved reviewed of the items for possible causes of DIF by a panel consisting of four bilingual science teachers.

METHODOLOGY

The achievement tests in this study is the primary school sixth grade national level science tests for 2005. As the test was administered at the beginning of the school year (February 2006), the most appropriate sample is the grade 7 secondary school pupils. All items in the science test were developed in Malay and then translated into Chinese and Tamil. During the national level examination, pupils who have learned science in the Malay language sat for the Malay version of the science test while pupils in the Chinese primary schools sat for the Chinese version and Indian pupils in the Tamil primary schools sat for the Tamil version. Besides the three language versions of the test, each version f the science test was accompanied with the English version as well. This practice started in 2003 with the implementation of teaching science in English. This step is adopted to help pupils who are not proficient in the English language so as not to be disadvantaged in assessment of science and mathematics achievement.

This study dealt only with science test in the Malay and Chinese language. The Tamil version was not compared due to the small number of Indian students in the schools chosen in the study. The science achievement test in the Primary School Assessment Test contained 30 multiple-choice items with four options given. All items have either diagrams or pictures to aid in explaining the questions.

Two versions of the Primary School Science Assessment Test were used. The Malay version was administered to 424 Malay students who received science instruction in Malay. Another 400 seventh grade students in the national secondary schools who received science instruction in Chinese were administered the Chinese version of the science achievement test.

STATISTICAL ANALYSIS

The assumption that multiple language test forms developed by a group of testing specialist and bilingual experts will measure comparable constructs need to be verified empirically (Ercikan et. al., 2004).

Multidimensional DIF Analysis

Multidimensional Model for DIF (MMD) is based on the assumption that multidimensionality produces DIF. A dimension is a substantive characteristic of an item that can affect the probability of a correct response. The main construct that the test intended to measure is the primary dimension. Besides the primary dimension, DIF items measure addition dimension that produce DIF (Ackerman, 1992; Roussos & Stout, 1996; Shealy & Stout, 1993). The addition dimensions are referred to as the secondary dimensions. When primary and secondary dimensions characterize responses, the data are considered multidimensional. Secondary dimensions may be part of the test construct being assessed intentionally or unintentionally. The Roussos-Stout DIF analysis paradigm is built on the foundation provided by MMD. The DIF hypothesis specifies whether an item designed to measure the primary dimension also measure a secondary dimension, thereby producing DIF (Gierl & Khaliq, 2001).

The simultaneous item bias test (SIBTEST), a nonparametric statistical method of assessing DIF in an item or bundle of item is used. This method based on Shealy-Stout's (1993) multidimensional model for DIF with the basic assumption that multidimensionality produces DIF. SIBTEST detects DIF by comparing the responses of examinees in the reference and focal groups that have been allocated to the same group using their score on a "matching subtest" (Stout & Roussos, 1995).

The magnitude of item DIF is interpreted using the general guidelines provided by Roussos and Stout (1996):

(a) neligible or A-level DIF: Null hypothesis is rejected and $|\hat{\beta}_{uni}| < 0.059$;

(b) moderate or B-level DIF; Null hypothesis is rejected and $0.059 \leq |\hat{\beta}_{uni}| < 0.088$; and

(c) large or C-level DIF; Null hypothesis is rejected and $|\hat{\beta}_{uni}| \geq 0.088$.

Item parameter estimation

With the computer program BILOG-MG V3.0, a 1-parameter logistic model will be used to estimate the item parameters. The model assumes that the item discrimination parameters are equal across the two groups being compared. Essentially, the differences in item difficulty parameter is assessed to account for group differences that cannot be explained by the test impact. Differences across two groups of examinees in item difficulty means that the item is more difficult for one group relative to the other group of examinees.

RESULTS

Psychometric Characteristics of the Test

The psychometric characteristics of the two versions of the science achievement test are presented in Table 1.

Table 1
Descriptive characteristic for Malay and Chinese Science Test

Characteristic	Malay	Chinese
No. of examinees	424	400
No. of items	30	30
Mean	23.67	21.44
Standard deviation	3.14	5.15
Skewness	-1.87	-1.21
Kurtosis	6.20	1.17
Internal consistency[a]	0.71	0.84
Mean item difficulty	-2.55	-1.59
SD item difficulty	2.27	1.70
Range of item difficulty	-6.70 – 3.09	-5.343 – 1.96

[a]Cronbach's alpha

The Malay examinees seem to perform better than the Chinese examinees. The mean score for the Malays is 23.67, SD=3.14 compare to the Chinese mean score of 21.44 and SD=5.15. The distribution of the Malays was more negatively skewed (-1.87) than the distribution of the Chinese(-1.21). The distribution of the Malays peak higher (6.20) than the distribution of the Chinese (1.17). Nonetheless, the internal consistency of the test items was slightly higher for the Chinese group (0.84) even though the items were shown to be easier for the Malay examinees (0.71).

SIBTEST yields two statistics of interest: the p-value and the Beta estimates that describe the size of the difference. An initial DIF analysis was run in which each item was screened using all of the remaining items as a matching subtest. The item(s) with the highest Beta estimate was "ignored" and the automatic analysis repeated. Successive iterations of the process eventually identified a subset of item that exhibited no statistically significant DIF. The result is shown in Table 2.

Table 2
Results from SIBTEST that screened each item

Item	Beta	p-value	Class	Favoured	ρ	r
1	0.000	0.974	A		0.966	0.348
2	0.103	0.000*	C	Malay	0.850	0.542
3	-0.262	0.000*	C	Chinese	0.328	0.166
4	-0.039	0.022	A		0.869	0.454
5	0.000	0.969	A		0.953	0.384
6	-0.054	0.123	A		0.681	0.393
7	0.012	0.522	A		0.902	0.522
8	0.348	0.000*	C	Malay	0.353	0.172
9	-0.035	0.022	A		0.925	0.415
10	-0.080	0.018*	B	Chinese	0.732	0.472
11	-0.103	0.000*	C	Chinese	0.842	0.478
12	0.0704	0.040*	B	Malay	0.617	0.303
13	-0.008	0.742	A		0.831	0.419
14	0.182	0.000*	C	Malay	0.642	0.517
15	-0.101	0.001*	C	Chinese	0.836	0.456
16	0.110	0.000*	C	Malay	0.809	0.551
17	-0.008	0.372	A		0.973	0.475
18	0.113	0.000*	C	Malay	0.826	0.481
19	0.117	0.000*	C	Malay	0.777	0.550
20	-0.027	0.274	A		0.828	0.487
21	-0.332	0.000*	C	Chinese	0.422	0.291
22	0.009	0.633	A		0.893	0.550
23	0.056	0.008	A		0.876	0.519
24	0.008	0.765	A		0.817	0.447
25	0.009	0.708	A		0.106	-.105
26	0.009	0.526	A		0.936	0.441
27	0.098	0.000*	C	Malay	0.864	0.438
28	0.100	0.000*	C	Malay	0.791	0.522
29	-0.003	0.839	A		0.941	0.439
30	-0.430	0.000*	C	Chinese	0.404	0.167

*= significance at $p<.05$

ρ : Proportion correct response on the item

r : Point biserial

Using a critical p-value of .05, 13 items exhibit large DIF. These are items 2, 3, 8, 11, 14, 15, 16, 18, 19, 21, 27, 28 and 30. While two items, items 10 and 12 exhibit moderate DIF. Using items with large DIF (Item 2, 3, 8, 11, 14, 15, 16, 18, 19, 21, 27, 28 and 30) as the suspect subtest, SIBTEST analysis was run again. The Beta estimate of 0.018 was not significant. There was no difference in difficulty between the two groups.

BILOG is used to calibrate item parameters so as to identify item(s) that may function differentially between the comparison groups. The two groups are Group1, referenced group in

SIBTEST; and Group 2, focal group in SIBTEST. For each group, BILOG-MG outputs estimates and standard errors for item difficulty and discrimination, the latter being equivalent across groups (Appendix A). The metric will be defined by setting the mean for Group 1 at 0 and the standard deviation at 1.0, whereas these values will be estimated in the Group 2 sample.

The mean and standard deviations for the group difficulty parameter estimates are shown in Table 1. The mean difficulty of the Chinese version is 0.956 above that of the Malays version. The adjusted value for the difficulty in group 2 is 0.956 (-2.548-(-1.592)). To create a set of item difficulty parameter contrasts, this value, 0.956 is subtracted from each difficulty value in Group 2. The contrast for the 30 items are shown in the third column of Table 3.

Table 3
Adjusted values for difficulty parameter and group difficulty differences

Item	Group 1	Group 2	Group (2-1)	SIB
1	-5.598	-5.196	0.402	0.698
2	-4.021	-2.515	1.507	4.459*
3	1.762	-0.526	-2.288	-11.556*
4	-3.030	-3.377	-0.347	-1.218
5	-4.658	-4.938	-0.280	-0.600
6	-1.165	-1.890	-0.725	-3.607
7	-3.956	-3.579	0.377	1.039
8	-0.140	1.429	1.569	7.729*
9	-3.781	-4.374	-0.593	-1.647
10	-1.515	-2.263	-0.748	-3.416*
11	-2.272	-3.490	-1.218	-4.445*
12	-1.010	-1.202	-0.192	-1.032
13	-2.557	-2.970	-0.413	-1.619
14	-1.834	-0.795	1.039	4.766*
15	-2.296	-3.322	-1.026	-3.916*
16	-3.298	-2.152	1.146	3.979*
17	-6.704	-5.343	1.361	1.634
18	-3.429	-2.358	1.071	3.706*
19	-2.924	-1.855	1.069	4.127*
20	-2.702	-2.788	-0.086	-0.328
21	1.150	-1.141	-2.290	-11.744*
22	-4.020	-3.350	0.0671	1.848
23	-4.021	-2.994	1.027	2.926*
24	-2.672	-2.617	0.055	0.218
25	3.090	1.958	-1.132	-4.087*
26	-4.870	-4.165	0.706	1.562
27	-3.958	-2.788	1.170	3.622*
28	-3.065	-1.993	1.072	4.030*
29	-4.658	-4.419	0.239	0.546
30	1.707	-1.434	-3.141	-15.705*

The SIB, Standardized index of Bias is the DIF contrast divided by the joint S.E. of the two DIF measures. Muraki and Engelhard (1989) noted that a criterion of about 2 may be used to judge an item to exhibit DIF. Based on this criteria, items 2, 3, 8, 10, 11, 14, 15, 16, 18, 19, 21, 23, 25, 27, 28, 30 exhibit DIF.

Comparison of the two statistical analyses

Items that were detected as DIF from SIBTEST and BILOG-MG were compared as in Table 4.

Table 4
Comparing that statistical analyses for items DIF detection

	Common items	Unique to the analysis
SIBTEST	2, 3, 8, 10, 11, 14, 15, 16, 18,	12 (moderate)
BILOG-MG	19, 21, 27, 28, 30	23, 25

The two statistical analyses have a high degree of congruence in detecting DIF in science items. 14 similar items were identified as functioning differently for the two groups by both SIBTEST and BILOG-MG. Item 12 was identified as moderately DIF by SIBTEST only and items 23 and 25 were identified as DIF by BILOG-MG..

Items reviewed by science teachers anticipated that different versions of the science achievement test contribute marginally to the DIF. It was argue that all 30 items use pictures and illustrations and the use of words and long sentences were rare. However, item 2 was identified as DIF due to the scientific terminology in Chinese which will provide clue to the correct answer. Item 3 was noted as not clear in the Chinese version which may results in the Chinese item become more difficult. Item 15 was identified as an easier item in Chinese as the terminologies used provided clue to the meaning of the phenomena of light that occurs. Item 19 was identified as more difficult in Chinese as terminologies used in describing changes in states of matter were almost identical.

CONCLUSION

Both statistical analyses identified 14 of 30 items on the primary science items displayed DIF. The results suggested substantial psychometric differences between the two language versions of the science test at the item level. Approximately, 50% (15 items by SIBTEST) to 53% (16 items by BILOG-MG) of the items were identified as DIF by both detection methods. These results reveal that a relatively large number of DIF items in the science achievement test. This finding is similar to those reported by other researchers in the area of test translation and adaptation (e.g. Allalouf, Hambleton, & Sireci, 1999, Gierl, Rogers, & Klinger, 1999; Gierl & Khaliq, 2001; Ercikan et. Al., 2004)

There is only a slight difference between the DIF detection patterns of SIBTEST and BILOG-MG. First, the BILOG-MG DIF detection method identified larger number of DIF items. Second, BILOG-MG identified more DIF items in favour of Malay examinees.

The accuracy of a translated test is crucial to ensure that both language versions of the test are measuring the same targeted ability. In large-scale testing situations, DIF is a constant concern as poorly translated items may put some students at a disadvantage (Hambleton, 1994; Hambleton & Patsula, 1998). The findings of this study highlight that comparability of different language versions of assessment cannot be assumed, and empirical examinations of comparability is essential to validity of interpretations. However, statistical outcomes alone cannot positively determine the cause of the difference. Substantive analysis help to investigate the multiple sources of incomparability that contribute to differences in constructs assessed. If the factors affecting the DIF of different language versions of items could be predicted, these could be taken into account in the test development process, thus resulting in improved decisions regarding test construction, scoring and equating. Results on sources of DIF can be used to develop guidelines and test construction principles for reducing DIF on translated tests.

The Standards (1999)recommended test developers should strive to identify and remove language, symbols, words and phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender or other groups, except when judged necessary for adequate representation of the domain. As such, there is a need to identify the sources of DIF due to translation so that these will be taken into consideration during test development to improve the test. If sources of translation DIF can be anticipated, then test developers could monitor test construction, translation, and adaptation practices to ensure the different language forms of the test are comparable across language groups. These will reduce the number of items that do not function equivalently across languages.

REFERENCE

Allalouf, A., Hambleton, R.K., & Sireci, S.G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement,* 36, 185-198.

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.

Budgell, G.R., Raju, N.S., & Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement,* 19, 309-321.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practices,* 17(1), 31-44.

Ercikan, K., Gierl, M.J., McGreith, T., Puhan, G. & Koh, K. (2004). Comparability of bilingual versions of assessment: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education,* 17(3), 301 -321

Angoff, W.H., & Cook, L.L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Rest: College Entrance Entrance Test. In T. Oakland & R.K. Hambleton (Eds.). *International perspectives on academic assessment* (pp. 207-217). Boston: Kluwer Academic

Ercikan, K. (1998). Translation effects I international assessment. International Journal of Educational Research, 29, 543-553.

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

Ercikan, K. and McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In D. Robitaille & A. Beaton (Eds.), Secondary aalysis of the TIMSS results: A synthesis of current research (pp. 391-407). Dordrecht, the Netherlands: Kluwer

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

Gierl, M.J., & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement,* 38, 164-187.

Gafni and Canaan-Yehoshafat (1993). An examination of differential item functioning for Hebrew and Russian-speaking examinees in Israel. Paper presented at the Conference of the Israeli Psychological Association, Ramat-Gan.

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

Hambleton, R.K. (1994). Guidelines for adapting educational psychological tests: A progress report. *European Journal of Psychological Assessment,* 10, 229-234

Hambleton, R.K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research.* 45, 153-171.

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment. Mahwah, NJ: Erlbaum.

Hulin, C.L. (1987). A psychometric theory of evaluations of item and test translations: Fidelity across languages. *Journal of Cross-cultural Psychology,* 67, 115-142.

Jöreskog, K. G. & Sörbom, D. (1993). LISREL 8.14: A computer program for structural equation modeling. Chicago, IL: Scientific Software

Muraki, E., & Engelhard, G. (1989). *Examining differential item functioning with BIMAIN.* Paper presented at the annual meting of the American Educational Research Asociation, San Francisco

Roussos, L.A., & Stout, W.F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenzel type I error performance. *Journal of Educational Measurement,* 33, 215-230

Shealy, R., & Stout, W.F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika,* 58, 159-194

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics,* 6, 317-375

Sireci, S.G., & Berberoglu, G. (2000). Using bilinguals to evaluate translated assessment questions. *Applied Measurement in Education,* 13(3), 229-248.

Sireci, S.G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations in international uses*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Standards for education and psychological testing. (1999). Washington, D.C.: American Education Research Association, American Psychological Association, and National Council on Measurement in Education

Van der Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Jouranl of Psychological Assessment.* 13, 21-29.

van der Vijver, F., & Tanzer, N.K. (1998). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, 47(4), 263-279.

COMPARING SCIENCE ACHIEVEMENT TESTS USING DIFFERENTIAL ITEM FUNCTIONING

By
Ong Saw Lan
osl@usm.my
Lee Lay Wah
Universiti Sains Malysia

Malaysian Education System

- Three type of primary schools: Malay, Chinese and Tamil
- National level achievement tests: Reading & writing, mathematics and science
- Mathematics and science assessed in Malay, Chinese & Tamil

Rationale

Research has shown that translating a test from a source language to a target language does not necessarily produce two psychometrically equivalent tests

(Allalouf, Hambleton, & Sireci, 1999; Budgell, Raju & Quartetti, 1995; Ercikan, 1998; Hambleton, 1993)

Methods of Comparison

- Judgmental Method
- DIF statistical analysis

Judgmental Method

Reviewers are generally poor at predicting which item would function differently across groups

(Engelhard, Hansce, & Rutledge, 1990; Gierl & McEwen, 1998)

DIF Statistical Analysis

- Procedure to identify items that function differently between different groups
- DIF occurs when item is substantially more difficult for one group than for another group after the group ability is taken into account
- Limitation: Do not reveal the causes of DIF

Question

1. How comparable are the two language versions of the science achievement test?
2. To what extent the two statistical analyses correspond each other in detecting DIF in science items?

Research Design

- Stage One: Detect DIF in science items using two different statistical analyses
- Stage two: four bilingual science teachers reviewed items for possible causes of DIF

Instrument

- Primary School Science Achievement Test in Malay & Chinese
- 30 multiple-choice items with four options
- All items have diagram or picture

Sample

- Grade seven students
- Malay version administered to 424 Malay students who received science instruction in Malay
- Chinese version administered to 400 Chinese students who received science instruction in Chinese

Statistical Analysis

- Multidimensional Model for DIF: multidimensionality produces DIF
- Besides the primary dimension, DIF items measure additional dimension that produces DIF

SIBTEST

- Simultaneous item bias test (SIBTEST) based on Shealy-Stout's (1993) multidimensional model for DIF
- Roussos and Stout (1996)
  - (a) $|\beta| < 0.059$ neligible or A-level DIF
  - (b) $0.059 \leq |\beta| < 0.088$ moderate or B-level DIF
  - (c) $|\beta| \geq 0.088$ large or C-level DIF

Item Parameter Estimation

- Difficulty parameter estimated for focal and reference group using BILOG-MG V3.0
- Differences in item difficulty parameter is assessed to account for group difference that cannot be explained by the test impact

Psychometric Characteristic

Characteristic	Malay	Chinese
No. of examinees	424	400
No. of items	30	30
Mean	23.67	21.44
S.D.	3.14	5.15
Skewness	-1.87	-1.21
Kurtosis	6.20	1.17
Internal Consistency	0.71	0.84
Mean Item Difficulty	-2.55	-1.59
S.D. Item Difficulty	2.27	1.70
Range of item difficulty	-6.70 to 3.09	-5.343 to 1.96

SIBTEST

Item	Beta	p-value	Class	Favoured	ρ	r
1	0.000	0.974	A		0.966	0.348
2	0.103	0.000*	C	Malay	0.850	0.542
3	-0.262	0.000*	C	Chinese	0.328	0.166
4	-0.039	0.022	A		0.869	0.454
5	0.000	0.969	A		0.953	0.384
6	-0.054	0.123	A		0.681	0.393
7	0.012	0.522	A		0.902	0.522
8	0.348	0.000*	C	Malay	0.353	0.172
9	-0.035	0.022	A		0.925	0.415
10	-0.080	0.018*	B	Chinese	0.732	0.472
11	-0.103	0.000*	C	Chinese	0.842	0.478
12	0.0704	0.048*	B	Malay	0.611	0.303
13	-0.008	0.742	A		0.831	0.419
14	0.182	0.000*	C	Malay	0.642	0.517
15	-0.101	0.001*	C	Chinese	0.836	0.436
16	0.110	0.000*	C	Malay	0.809	0.551
17	-0.008	0.372	A		0.973	0.475
18	0.113	0.000*	C	Malay	0.826	0.481
19	0.117	0.000*	C	Malay	0.777	0.550
20	-0.027	0.274	A		0.828	0.487
21	-0.332	0.000*	C	Chinese	0.422	0.291
22	0.009	0.633	A		0.893	0.550
23	0.056	0.008	A		0.876	0.519
24	0.008	0.765	A		0.817	0.447
25	0.009	0.708	A		0.106	-.103
26	0.009	0.526	A		0.936	0.441
27	0.098	0.000*	C	Malay	0.864	0.438
28	0.100	0.000*	C	Malay	0.791	0.522
29	-0.003	0.839	A		0.941	0.439
30	-0.430	0.000*	C	Chinese	0.404	0.167

Difficulty Parameter

Item	Group 1	Group 2	Group (2-1)	SIB
1	-5.598	-5.196	0.402	0.698
2	-4.021	-2.515	1.507	4.459*
3	1.762	-0.526	-2.298	-11.556*
4	-3.030	-3.377	-0.347	-1.218
5	-4.658	-4.938	-0.280	-0.600
6	-1.165	-1.890	-0.725	-3.607
7	-2.956	-3.579	0.377	1.039
8	-0.140	1.429	1.569	7.729*
9	-3.781	-4.374	-0.593	-1.617
10	-1.515	-2.263	-0.748	-3.416*
11	-2.272	-3.490	-1.218	-4.455*
12	-1.100	-1.292	-0.192	-1.632
13	-2.557	-2.970	-0.413	-1.619
14	-1.834	-0.795	1.039	4.766*
15	-2.296	-3.322	-1.026	-3.916*
16	-3.298	-2.152	1.146	3.979*
17	-6.704	-5.343	1.361	1.634
18	-3.429	-2.358	1.071	3.706*
19	-2.924	-1.855	1.069	4.127*
20	-2.702	-2.788	-0.086	-0.326
21	1.150	-1.141	-2.290	-11.744*
22	-4.020	-3.350	0.0671	1.848
23	-4.021	-2.994	1.027	2.926*
24	-2.672	-2.617	0.055	0.216
25	3.090	1.958	-1.132	-4.027*
26	-4.870	-4.165	0.706	1.562
27	-3.958	-2.788	1.170	3.622*
28	-3.065	-1.993	1.072	4.030*
29	-4.658	-4.419	0.239	0.546
30	1.707	-1.434	-3.141	-15.705*

Comparing Two analyses

	Common items	Unique to the analysis
SIBTEST	2, 3, 8, 10, 11, 14, 15, 16, 18,	12 (moderate)
BILOG-MG	19, 21, 27, 28, 30	23, 25

Items Reviewed

- Marginal DIF anticipated
- All 30 items use pictures and illustrations
- Item 2: Terminology in Chinese is easier
- Item 3: not clear in Chinese
- Item 15: Terminology provide clue
- Item 19: Terminologies identical

Conclusion

- 14 items displayed DIF by both statistical analysis
- Larger number of DIF items by BILOG-MG (15 items), 16 items by SIBTEST,
- Relatively large number of DIF items
- BILOG-MG identified more DIF items in favour of Malay examinees

Thank You