International Journal of Educational Excellence
(2020) Vol. 6, No. 2, 47-62
ISSN 2373-5929
DOI: 10.18562/IJEE.057

Studying the Equivalence of Two Language Versions of a Large Scale Assessment: A Comparison of Test Takers in United States and Puerto Rico

Jorge Carvajal Espinoza [a], & John Poggio [b]

Received: 1 October 2020 • Accepted: 7 November 2020

Abstract: This study collected evidence regarding the equivalence of items across language forms by applying DIF methodologies to both the cognitive and affective domains of a large scale assessment and explored the utility of the Liu-Agresti estimator of the cumulative common odds ratio for identifying polytomous DIF. The illustrated use of the Liu-Agresti estimator appears to be promising to the understanding of the phenomenon of polytomous DIF. Although the differential functioning of the polytomous items identified as large DIF could not be accounted for based on translation, it does not follow that there are not other causes for the apparent differential functioning. In particular, since these items tap into affect, behavior and attitudes, it could be cultural differences or impact that account for such differences.

Key-words: test takers; assessment; second-language test adaptation; translation DIF; Liu-Agresti estimator; large scale assessment.

1. Introduction

One approach to studying measurement equivalence is differential item functioning. As several states (Texas, California, Kansas and others) move to developing Spanish language assessments due to the increasing Latino/Hispanic population, assessing the equivalence and validity of second-language assessments are important considerations within large scale testing programs. Not only should the meaning of a test be consistent across persons

[a] Universidad de Costa Rica (UCR), Costa Rica. 0000-0003-0204-4894. [b] University of Kansas, United States. 0000-0001-9432-4871. Correspondence: Jorge Carvajal Espinoza, Ciudad Universitaria Rodrigo Facio Brenes, San José, San Pedro, Costa Rica. carvajalespinoza@ucr.ac.cr

within a cultural group, that meaning must be consistent across cultural groups (Van de Vijver and Poortinga, 1997). Second-language test adaptation presents challenges in certain domains. For example, in mathematics, the construct of interest may be focused on computation skills and the purpose of the test is to look for a demonstration of those skills. For this domain, the language in which the performance is assessed may be of little or no interest (Hambleton and Patsula, 1999). However, desired inferences in other content domains (science, for example) present questions pertaining to equivalence.

When a test is translated, equivalence of items across language forms is a critical issue to be considered (Price, 1998). The Differential Item Functioning (DIF) methodology is a family of techniques commonly used as a means to evaluate this equivalence (Sireci and Khaliq, 2002; Emenogu and Childs, 2003; Ulterwijk and Vallen, 2003; Sireci, Fitzgerald and Xing, 1998; Gierl, Rogers and Klinger, 1999; Robin, Sireci and Hambleton, 2003). Items function differently (DIF is said to exist) when test-takers of equal ability differ substantially, on average, according to their group membership in their responses to a given item (AERA, APA, NCME, 2014). Since item-level DIF may not manifest itself in scale-level analyses (Zumbo, 2003) it is important and primary to carry out analyses of equivalence at the item level.

The psychometric literature pertaining to equivalence across language forms beyond the cognitive domain (i.e., affective) is lacking. As an increasing number of educational, credentialing, and psychological tests are being adapted for use in other languages, a treatment of equivalence across different language forms within both the cognitive (test's of maximum performance) and affective (tests of typical behavior) domains in the context of a large scale assessment system is needed.

The purpose of the present study is twofold: to collect evidence regarding the equivalence of items across language forms by applying DIF methodologies to both the cognitive and affective domains of a large scale assessment as well as to explore the utility of the Liu-Agresti estimator of the of the cumulative common odds ratio (Liu and Agresti, 1996) for identifying polytomous DIF. The use of this estimator for polytomous DIF analysis was proposed by Penfield and Algina (2003). This estimator has not been employed to analyze polytomous DIF in the affective domain and it has only been applied to real data in a test of dichotomous and polytomous cognitive items by Penfield and Camilli (2007).

In addressing these purposes, a traditional Mantel Haenszel (MH) analysis is employed for one sample of cognitive, dichotomous items while for the affective, polytomous items the Liu-Agresti approach is applied across different samples and compared with the Logistic Discriminant Function Analysis (LDFA) and the Mantel for DIF detection procedures. A classification of polytomous DIF items using the Liu-Agresti estimator, which is comparable to the ETS DIF classification scheme for the dichotomous case, is also explored.

The vehicle for the current analysis is a large scale assessment measuring cognitive achievement for planned instructional objectives and related specific affective outcomes associated with those objectives. As an integrated assessment tool, this test provides knowledge questions (cognitive domain) and questions related to beliefs, attitudes, practices, and perceptions (affective domain). As a multilevel, age appropriate tool, this test is used in grade 5 (Level 1); grades 8 or 9 (Level 2); and grades 11 or 12 (Level 3). The present study evaluates the equivalence of the Level 2 Spanish version of the test.

The Spanish version of this form was made available for the first time in the year 2005 for the target population of English Language Learner students (EL) in the US whose native language is Spanish. As this was the first year this form was available, relatively few ELL students took this version in the 2005 administration. Nonetheless, 61 native Spanish speaking 8th grade students in Puerto Rico that took the Spanish version in 2005 were identified. This group represents the focal group in this study. Approximately 66,000 8th graders took the English version of the test in the US in 2005 and various samples from this population constitute the reference groups.

Although DIF studies based on small sample sizes can be problematic (Fidalgo, Ferreres, and Muñiz, 2004) it is not an uncommon situation for testing settings to encounter such small samples: "… state boards, certification and licensure agencies, and others, often make contractual requirements for DIF analysis, regardless of the statistical appropriateness of the sample size" (Parshall and Miller, 1995, p. 314). In the current study, the high reliability of scores from the analyzed test as well as the thick matching scheme used are factors that somewhat mitigate problems inherent to small samples. (Zwick, Thayer and Mazzeo, 1997; Donoghue and Allen, 1993; Clauser, Mazor and Hambleton, 1994)

2. Description of DIF procedures used in the study

2.1. Mantel Haenszel

The Mantel Haenszel (MH) procedure for detecting DIF in dichotomous items is widely used, including in situations where sample sizes are small. We denote the MH estimator of the common odds ratio by $\hat{\alpha}_{MH}$ and its logarithm by $\log\hat{\alpha}_{MH}$ (log odds ratio estimator). The MH log odds ratio in delta metric (Dorans and Holland, 1993) is denoted by MH D-DIF. The ETS classification flags dichotomous DIF items as Type C (large DIF) when $|\text{MH D - DIF}|$ is greater than 1.5 and statistically greater than 1 (Dorans &

Holland, 1993). Because MH D-DIF = -2.35$\left|\log \hat{\alpha}_{MH}\right|$, this classification is equivalent to saying that a DIF item is classified as Type C if $\left|\log \hat{\alpha}_{MH}\right|$ is greater than 0.64 and statistically greater than 0.43. In the current study, dichotomous Type C items are identified and they are considered for judgmental review.

2.2. Mantel for polytomous items

The Mantel for polytomous items is a generalization of MH for the dichotomous case (Zwick, Donoghue, and Grima, 1993). There is a chi-square test with 1 degree of freedom associated with this procedure. If the null hypothesis is rejected, the item is identified as evidencing DIF (Wang and Su, 2004). This approach takes into account the ordinal nature of categorical responses of polytomous items, making its application appropriate for the current study.

2.3. Logistic Discriminant Function Analysis (LDFA)

LDFA is a variation of logistic regression procedures proposed for polytomous DIF analysis by Miller and Spray (1993). Under this technique (Su and Wang, 2005)

$$\ln\left|\frac{P(G=1)}{P(G=0)}\right| = \alpha_0 + \alpha_1 X + \alpha_2 U + \alpha_3 XU$$

where U is the item score, G is the group indicator and X is the conditional total score. Under this framework (Kristjansson, Aylesworth, and McDowell, 2005), three equations are derived: an equation predicting group membership from X (model 1), an equation predicting group membership from X and U (model 2), and the equation shown above predicting group membership from X, U, and their interaction (model 3). Based on the computation of a likelihood ratio goodness-of-fit statistic G^2, a significant result in the comparison between G^2 in model 3 and model 2 is evidence of the existence of non uniform DIF (a statistically significant interaction exists), whereas only a significant result in the comparison of G^2 between model 2 and model 1 is evidence of the presence of uniform DIF (i.e., no significant interaction).

2.4. The Liu-Agresti Estimator

This is an estimator of the common odds ratio across response categories of an ordinal response variable (Liu and Agresti, 1996). Penfield and Algina (2003) present the various formulas for this estimator ($\hat{\alpha}_{LA}$) as well as its properties, which are summarized below.

$\hat{\alpha}_{LA}$ is a generalization of $\hat{\alpha}_{MH}$, when the number of category responses is 2, $\hat{\alpha}_{LA}$ reduces to $\hat{\alpha}_{MH}$. Similarly to $\log \hat{\alpha}_{MH}$, $\log \hat{\alpha}_{LA} = 0$ suggests no DIF is present, $\log \hat{\alpha}_{LA} > 0$ suggests DIF in favor of the reference group and $\log \hat{\alpha}_{LA} < 0$ suggests DIF in favor of the focal group.

The formula for $Var \log \hat{\alpha}_{LA}$ can be also found in the Penfield and Algina (2003) paper. These researchers state that since the common log odds ratio are asymptotically normally distributed, the statistic

$$z_{LA} = \frac{\log \hat{\alpha}_{LA} - C}{\sqrt{Var \log \hat{\alpha}_{LA}}}$$

can be used to test directional and nondirectional hypotheses concerning the value of the population cumulative common odds ratio, where C is a constant.

Penfield and Algina (2003) discuss the advantages of $\hat{\alpha}_{LA}$ over other statistics for detecting DIF in polytomous items. The similarity of $\hat{\alpha}_{LA}$ to $\hat{\alpha}_{MH}$ permits $\hat{\alpha}_{LA}$ to be used in approaches to DIF detection in polytomous items in a manner analogous to $\hat{\alpha}_{MH}$ in the dichotomous case; for example, using a combination of the magnitude of $\hat{\alpha}_{LA}$ along with the proper value of z_{LA} to assess degrees of DIF in similar manner to the ETS dichotomous item classification scheme. According to these researchers, other possible applications include Bayesian approaches to investigate the probability that the polytomous items have varying levels of DIF and examining the presence of differential test functioning in tests comprised of polytomous items.

3. Method

For the translation of the test, two translators independently conducted a translation into Spanish. Following that procedure, a consensual validation of the translation was performed. A third translator then compared the English and consensual versions and offered edits and suggestions. The original two translators then prepared a unified version based upon those suggestions. Two Spanish native speakers who are experts in substantive field measured by the test reviewed the unified version and made final suggestions for editing and revisions. Their observations were incorporated and the final version was produced.

The Level 2 of this test consists of two parts. The first part contains 57 cognitive items that assorted cognitive knowledge. The assessment second part (affective) contains 46 questions the first 33 of which are 4-response category Likert items measuring various attitudes, thoughts, beliefs, and values directly related to and derived from the knowledge objectives measured by Part 1 cognitive knowledge of the test. The remaining 13 items measure perceptions of how frequently certain situations occur at school. As the two parts of the test measure distinct cognitive and affective components, a single total composite score was of no interest in the current study.

For the polytomous analysis, only the section 1 of 33 affective items were considered. The 13 perception and documentation items were not included for analyses as this collection of survey questions taps the perceived frequency of certain events and behaviors rather than an individual's attitudes or beliefs, clearly a different construct. Beyond relevance, given our small sample, it further would not be prudent to analyze the 13 perception items separately as it has been shown that short tests produce greater instability in DIF analyses.

Accordingly, the Part 1 cognitive matching variable (total score) for the dichotomous case ranges 0 to 57 and the matching variable for the polytomous analysis ranges 0 to 99 as each item was coded 0 to 3.

The focal group consisting of 61 Puerto Rican 8th grade students remained constant across all replications of the study in the dichotomous and polytomous analyses. The various samples used for the reference group were all randomly selected from the population of about 66,000 US 8th grade students who completed the English language test form.

For the dichotomous (cognitive) study, a sample of 348 US students was selected at random. For the polytomous (affective) study, samples of 350, 500, 1000 and 1500 were drawn to evaluate the behavior of $\hat{\alpha}_{LA}$ under varying sample sizes and to compare these results with the MH and LDFA analyses.

DIF analyses, except those corresponding to LDFA, were conducted using the DIFAS 2.0 software (Penfield, 2005). For the dichotomous case, DIFAS provides a classification based on the ETS scheme (type A, B, and C). For the polytomous analysis DIFAS 2.0 provides the Mantel test statistic, $\log\hat{\alpha}_{LA}$, and $Var\log\hat{\alpha}_{LA}$. LDFA analyses were conducted in SPSS. All statistical tests were evaluated at α = .05.

A particularly important decision in the current study regarding the analysis should be noted. The DIFAS 2.0 software package allows the user to choose the size of the interval (stratum) for which the statistics are computed. By default, the interval size is set at one. Thus, for example in the dichotomous case there would be by default 58 intervals given that the section contains 57 items. If intervals are set too wide, the impact of the test (difference in total score between the focal and reference group) could be a confounding of the DIF statistics. In the extreme case that only one interval exists for the whole test there would be a complete confounding. We used thick matching (Donoghue and Allen, 1993; Clauser et al., 1994), in an attempt to mitigate sparseness of data in the focal group, employing a reasonable size for the intervals of five (5). This decision was verified by doing the following: In the dichotomous case whereby logistic regression can be used for DIF analysis (Swaminathan and Rogers, 1990), under non uniform DIF in the model

$$\ln\left|\frac{P(U=1)}{P(U=0)}\right| = \beta_0 + \beta_1 X + \beta_2 G$$

where U is the item score, X is the conditional total score and G is the group with G=0 focal and G=1 reference, the value of β_2 should be close to that of the $\log \hat{\alpha}_{MH}$ (Penfield & Camilli, in press). Several samples were extracted for the purpose of comparing the difference between these two values when using intervals of size one (1) versus using intervals of size five (5). In this evaluation, not only was the mean of the difference closer to zero in the size five scenario, but the SD of the difference decreased in this case (for example, the mean changed from .029 for size one to .013 for size five; the SD decreased from .116 to .073 respectively), an indication that the estimation of $\log \hat{\alpha}_{MH}$ is at least as precise for the interval five case for our data. Thus grouping error was not an issue in this study.

For the same purpose of mitigating sparseness of data in the focal group, interval size five was used for the polytomous DIF analysis in DIFAS 2.0.

To date, no classification rule to assess the severity of DIF has been proposed in the literature using $\hat{\alpha}_{LA}$ for polytomous items. Penfield (personal communication, 2010) suggests that items exhibiting DIF can be identified using $\hat{\alpha}_{LA}$ and its properties by the combined criterion: $|\log \hat{\alpha}_{LA}|$ greater than about 0.6 and $\log \alpha_{LA}$ significantly different than zero. This combined criterion is similar to the ETS classification rule for dichotomous items and would include what in the dichotomous case are type B and C items without differentiating among them. Since the methodological decision was made for this study that only type C items in the dichotomous case would be identified and selected for judgmental analysis (i.e. items that exhibit greater DIF) a similar methodological decision was made for the polytomous case. Therefore, the decision was made to identify polytomous items as large DIF items if $|\log \hat{\alpha}_{LA}|$ is greater than 0.64 and $|\log \alpha_{LA}|$ is statistically greater than 0.43 in an analogy to the dichotomous case under the ETS classification. We will refer to these items as polytomous Type C items.Of interest in the current study is whether polytomous Type C items flagged by this criterion are also flagged as DIF items by the Mantel or LDFA approaches in order to collect evidence about the validity of such classification.

The logical review judgmental analysis of the Type C items (dichotomous and polytomous) was conducted with a panel of 3 bilingual native Spanish speakers, one of them a Puerto Rican. In addition, a teacher of the Puerto Rican examinees provided feedback, input and reaction regarding potential reasons for differential functioning of the type C items.

4. Results

4.1. Part A: Cognitive Dichotomous Scored Items

Part 1 Descriptive Statistics

Group	Subjects	Mean Score*	Score SD	Items mean	Alpha
PR	61	35.2	8.9	0.62	0.88
US	348	41.6	9.3	0.73	0.91

Total score: 57

Table 1. Descriptive Statistics (I)

Table 1 shows the mean for the Puerto Rican (Spanish version of the test) and US (English version) student samples for the cognitive test. The mean for the Spanish version is 35.2 while the mean on the English version is 41.6, with standard deviations of 8.9 and 9.3, respectively. The mean total score for the US group is 0.66 of a pooled SD higher that the mean total score for the Puerto Rican students, not an uncommon result in studies of this nature.

Table 1 also shows that mean item difficulty for US students is .73. In the Puerto Rican sample, the mean item difficulty is .62. Reliability coefficients as indexed by Cronbach’s Alpha for the US and PR samples are .91and .88, respectively.

Based on the results, 5 items were flagged as exhibiting large DIF (Type C): item 7, 13, 14, 24, and 47. Four of the 5 items favored the US student group. Item 13 favored the Puerto Rican student group. A judgmental analysis of these five items was conducted. In four of these items some translation issues were identified that might account for their apparent differential functioning. Because these items are operational and thus secure, in order to maintain the security of the test we do not report the full judgmental analysis.

4.2. Part B: Polytomous Affective Items

Part 2 Descriptive Statistics

Group	Subjects	Mean Score*	Score SD	Items mean**	Alpha
PR	61	65	12.8	1.97	0.88
US 350	350	68.7	13.1	2.08	0.91
US 500	500	67.9	12.3	2.06	0.9
US 1000	1000	67.6	12.8	2.05	0.91
US 1500	1500	67.8	12.6	2.05	0.91

***Total score**: 99

****Items coded 0 to3**

Table 2. Descriptive Statistics (II)

Table 2 shows that the mean score for the Puerto Rican group of students is 65.0 whereas the mean score for the four samples of US students ranges from 67.6 to 68.7. The score SD for the PR students is 12.8. The SDs of the US samples range from 12.3 to 13.1. The difference between the mean score of the Puerto Rican students and the 4US samples ranges from 0.20 to 0.28 of a SD (pooled for each sample). It is evident that for the polytomous case there is less of a difference between total score means between Puerto Rican and US students than in the case of the dichotomous cognitive maximum performance items.

Table 2 also shows the reliability coefficients. For the Puerto Rican sample the reliability is .88 whereas the reliability coefficient for the 4 US samples is.90 for the second sample and 0.91 for the other three. Additionally, Table 2 provides the item means for the different samples: 1.97 for the Puerto Rican group and between 2.05 and 2.08 for the US samples.

DIF results are presented in Table 3 for the three methods across the four different samples. The items classified as large DIF (Type C) in at least one of the samples using the implemented criterion for the Liu-Agresti estimator are included in this table.

Four items were classified as Type C across the four samples: items 4, 11, 15 and 33. Item 22 was classified as Type C in all samples except the 350-61 sample comparison. Item 9 was classified as Type C in two of the samples while being “borderline” in the others. In this particular case, it means that whereas $\left|\log\hat{\alpha}_{LA}\right|$ is greater than 0.64, statistical significance for $\left|\log\alpha_{LA}\right|$ greater than 0.43 in these two samples was not attained, the correspondent test statistics being close to the critical value. For items 4, 9, 11, 15, 22 and 33 there is a high degree of classification consistency despite the fact that the 4 sample sizes are distinct. On the other hand,item 19 was flagged in only one of the samples (350-61). This rate of flagging questions for DIF using .05 as the trigger, suggests a frequency of occasions nearing a chance Type 1 error rate. No other items among the 33 polytomous items were flagged as Type C in any sample, which also provides evidence about the consistency of the Liu-Agresti estimator across samples.

Polytomous DIF Items

Item		Sample 350-61	Sample 500-61	Sample 1000-61	Sample 1500-61
4	Liu-A	***	***	***	***
	Log odds ratio Lui-A	1.109	1.022	0.902	0.854
	Mantel	*	*	*	*
	LDFA		*	*	*
9	Liu-A	***	borderline	***	borderline
	Log odds ratio Lui-A	-0.998	-0.871	-0.9	-0.843
	Mantel	*	*	*	*
	LDFA		*	*	*
11	Liu-A	***	***	***	***
	Log odds ratio Lui-A	1.476	1.402	1.327	1.505
	Mantel	*	*	*	*
	LDFA		*	*	*
15	Liu-A	***	***	***	***
	Log odds ratio Lui-A	1.081	1.391	1.313	1.371
	Mantel	*	*	*	*
	LDFA		*	*	*
19	Liu-A	***			
	Log odds ratio Lui-A	-1.021	-0.799	-0.782	-0.785
	Mantel	*	*	*	*
	LDFA	*	*	*	*
22	Liu-A		***	***	***
	Log odds ratio Lui-A	-0.937	-1.041	-1.034	-0.996
	Mantel	*	*	*	*
	LDFA		*	*	*
33	Liu-A	***	***	***	***
	Log odds ratio Lui-A	-1.606	-1.696	-1.491	-1.586
	Mantel	*	*	*	*
	LDFA	*	*	*	*

*** Type C (Large DIF item)

* DIF item

Table 3. Polytomous DIF Items

Table 3 also presents the cumulative log odds ratio estimator ($\log \hat{\alpha}_{LA}$), which is a measure of effect size. It should be noted that $\log \hat{\alpha}_{LA}$ appears to be rather stable across samples, except for item 19. The sign of $\log \hat{\alpha}_{LA}$ on the table denotes the direction of the DIF. Thus, items 4, 11, and 15 show higher endorsement for students in the reference group whereas items 9, 19, 22, and 33 show higher endorsement for students in the focal group, after controlling for the construct measured by the affective items.

Further, Table 3indicates if items classified as Type C were also flagged as DIF items by the Mantel or the LDFA procedures. Note that any item classified as Type C was also flagged as DIF item by the Mantel procedure.

For the LDFA, a similar situation occurs in the three largest samples: all Type C items are also flagged by LDFA. The condition is different for LDFA with the 350-61 sample. In this case LDFA flagged only 2 of the 6 Type C items identified in this sample. It might be that LDFA does not have enough power for smaller samples.

The seven Type C items (4, 9, 11, 19, 15, 22 and 33) were reviewed by the panel for the purpose of a judgmental analysis. Contrary to the dichotomous (cognitive) case, no compelling evidence was uncovered regarding there being an issue with the translation. If in fact these items do exhibit large DIF (as opposed to the cause being Type I error), it would appear that the basis of the differential functioning is not related to the way the translation was made.

5. Discussion

With the increasing Latino/Hispanic population sitting for assessments in the US as well as the increasing use of translated tests, it is important to collect evidence of equivalence across languages. Understanding possible causes of differential item functioning is one of the advantages of carrying out translation DIF studies. In the current study it was possible to identify possible translation issues for four of the five large DIF items in the dichotomous case. This situation illustrates the importance of conducting such studies.

Traditionally, studies such as this have explored and examined DIF in the context of cognitive (dichotomous, i.e., right-wrong) assessment. DIF, in the context of polytomous assessment, has not received the same treatment in the literature. In addition to a traditional dichotomous analysis, this study explored DIF detection in polytomous items by employing an odds ratio estimator, the Liu-Agresti,which has not been reported in the literature for affective questions and that has only once applied to a real data set (Penfield and Camilli, 2007). Penfield and Algina (2003) offer that one of the possible advantages of this estimator is the implementation of a classification rule similar to the ETS classification that uses the MH for dichotomous items. An example of such a classification for polytomous items was implemented and illustrated in this study and compared with two more procedures, the Mantel and the LDFA for polytomous items. The behavior of the Liu –Agresti estimator across different sample sizes was also studied. When an item was classified as showing large DIF (Type C) by the criterion implemented in this study,it was also flagged as a DIF item by the Mantel and the LDFA (with the exception of a smaller sample for this latter method). However the opposite is not true. Though not reported in detail here, several items that were flagged by MH or LDFA were not classified as Type C items. It was also noted that $\log \hat{\alpha}_{LA}$, an effect size, appears to be rather stable across different combinations of sample sizes, as extreme as 61 for the focal and 1500 for the reference group. This is of particular importance because frequently in testing situations the sizes of the samples for focal groups can be rather low.

One cannot be certain that the items identified as Type C are in fact items with large DIF or whether they witness the occasion of and have been subject to Type I errors. However, that most of the items were consistently identified across different samples of different sizes and that they were mostly flagged by the other 2 procedures some evidence of a correct identification. Further treatment, possibly in the form of simulation studies, is needed to study more in-depth the behavior of this estimator.

Although the differential functioning of the polytomous items identified as Type C could not be accounted for on the basis of translation, it does not follow that there are not other causes for the apparent differential functioning. In particular, since these items tap into affect, behavior and attitudes, it could be cultural differences or impact that account for such differences. This is an area for further research.

From a methodological perspective, the illustrated use of the Liu-Agresti estimator appears to be promising and future research that focuses on the performance of this estimator not only in the direction of classification schemes but in other applications will help clarify the contribution of this estimator to the understanding of the phenomenon of polytomous DIF.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*, *16*, 55-73. doi: 10.1207/S15324818AME1601_3

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*, 185-198. doi: 10.1111/j.1745-3984.1999.tb00553.x

Clauser, B. Mazor, K, & Hambleton, R. (1994). The effects of score group width on the Mantel-Haenszel Procedure. *Journal of Educational Measurement, 31,* 67-68.

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

Donoghue, J. R., & Allen, N. (1993). Thin versus thick matching in the Mantel Haenszel Procedure for Detecting DIF. *Journal of Educational Statistics, 18*, 131-154. doi: 10.2307/1165084

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland and H. Wainer (Eds.), D*ifferential item Functioning* (pp.35-66). Hillsdale, NJ: Lawrence Erlbaum.

Duncan, T., Parent, L., Chen, L., Ferrara, S., & Johnson, E. (2002). *Study of a Dual Language Test Booklet in 8th grade Mathematics.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*, 199-215. doi: 10.1080/15305058.2002.9669493

Emenogu, B., & Childs, R. (2003). *Curriculum and Translation Differential item Functioning: A Comparison of Two DIF Detection Techniques.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*, 925-936. doi: 10.1177/0013164404267288

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

Hambleton, R.K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1-30

Hambleton, R., & Patsula, L. (2000). *Adapting Tests for use in Multiple languages and Cultures.* (Laboratory of Psychometric and Evaluative Research, Report No. 304). Amherst: University of Massachusetts, School of Education.

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

Kristjansson, E., Aylesworth, R., & McDowell, I. A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953. doi: 10.1177/0013164405275668

Liu, I-M, & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89-114. doi: 10.1177/026553220 101800104

Parshall, C.G., & Miller, T.R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of the performance under small-sample conditions. *Journal of Educational Measurement, 32*, 302-316.

Penfield, R. D. (2010). *Personal Communication*. doi: 10.1111/j.1745-3984.1995.tb00469.x

Penfield, R. D. (2005). Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151. doi: 10.1177/0146 621603260686

Penfield, R. D. & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp.125-167). New York: Elsevier..

Penfield, R. D. & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 343-370. doi: 10.1111/j.1745 -3984.2003.tb01151.x

Price, L. (1999). D*ifferential Functioning of Items and Tests Versus the Mantel-Haenszel Technique for Detecting Differential item Functioning in a Translated test.* Paper presented at the meeting of the American Alliance of Health, Physical Education, Recreation, and Dance, Boston, MA.

Price, L., & Oshima, T. (1998). *Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Miller, T., & Spray, J. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement, 30*, 107-122.

Muñiz, J., & Hambleton, R. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115-135. doi: 10.1207/S15327574IJT0102_2

Robin, F., Sireci, S., & Hambleton, R. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*, 1-20. doi:10.1207/S15327574IJT0301_1

Sireci, S., & Khaliq, S. (2002). *An Analysis of the Psychometric Properties of Dual Language Test Forms*. (Center for Educational Assessment, Report No. 458). Amherst: University of Massachusetts, School of Education.

Sireci, S., Fitzgerald, C., & Xing, D. (1998). *Adapting Credentialing Examinations for International Uses.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

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

Su, Y., & Wang, W. (2005). Eficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function Analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*, 313-350. doi: 10.1207/s15324818 ame1804_1

Ulterwijk, H., & Vallen, T. (2003). Test bias and differential item functioning: a study of the suitability of the CITO primary education final test for second generation immigrant students in the Netherlands. *Studies in Educational Evaluation, 29*, 129-143.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37.

Wang, W. & Su, Y. (2004). Factors influencing the Mantel and Generalized Mantel-Haenszel methods for the assessment of differential item functioning in Polytomous items. *Applied Psychological Measurement, 28*, 450-480. doi: 10.1177/0146621604269792

Zenisky, A., Hambleton, R., &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*, 51-64.

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing 20*, 136-147. doi: 10.1191/0265532203lt248oa

Zwick, R., Thayer, D. & Mazzeo, J. (1997). *Describing and Categorizing DIF in Polytomous Items*. ETS Research Report 97-05.

Zwick, R., Donoghue, J. & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.