Language Cert

Potential Bias in LanguageCert IESOL Items: A Differential Item Functioning Analysis

David Coniam
and
Tony Lee

Abstract

Differential Item Functioning (DIF) analysis is a statistical procedure undertaken to explore whether any subgroup of test takers sitting a test or exam is being unfairly disadvantaged or indeed advantaged. Investigating DIF is key to understanding and dealing with test bias, a necessary though complex requirement of leading test providers. To date, PeopleCert has not been in a position to address this issue in any depth. This has the potential to diminish the organization's standing in the international assessment community. However, preliminary work on test bias and DIF has now begun and the methodology required to carry it out has been identified and is described in this paper.

The current study reports on a DIF analysis of the six IESOL exams aligned to the CEFR and delivered by LanguageCert between 2018-2020. For each CEFR level. Four variables, namely mother tongue, age, gender and test centre were explored. DIF analysis was conducted using the computer program Winsteps, with DIF strength reported in line with Zwick (1999).

Some moderate-to-large DIF was reported for *mother tongue* and *decade of birth (*a recording of *age*). This, however, may well be due to the fact that these two categories are very diverse with only very few entries.

For *gender* – typically a key variable in the exploration of DIF – a very low incidence of 3% DIF was reported. For *centre* (i.e., a comparison of OLP vs non-OLP delivery), zero and moderate-to-large DIF was observed. An examination of reading or listening items indicated that there was no predominance of DIF in either skill.

These are encouraging preliminary findings and confirm that the six *LanguageCert* tests analysed here are showing relatively low levels of bias. Mechanisms will need to be put in place to monitor DIF on all LanguageCert products going forward. This will include gathering more information about candidates in a systematic and comprehensive manner so that important potential sources of bias can be investigated.

Background

Differential Item Functioning (DIF) analysis is a statistical procedure undertaken to explore whether any subgroup of test takers sitting a test is being unfairly disadvantaged. The exploration of potential sources of bias among subgroup types typically investigate variables such as gender, cultural affiliation, age etc. Indeed, in many DIF studies, and often for political reasons, key variables studied have tended to be gender and ethnicity (Ferne & Rupp, 2007).

Evidence of DIF may be apparent in a test item (or indeed on an entire test) when the responses of two groups of test takers who should have equal "latent trait ability" show different probabilities in terms of correctly answering a test item (Swaminathan & Rogers, 1990). While DIF analysis has been used for a considerable time by the general educational community, it is only in the past couple of decades that the use of DIF has become more prevalent in the language assessment field (Aryadoust et al., 2011; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000). Ferne & Rupp (2007) provide a cogent synthesis of 15 years of research on DIF in language testing.

While DIF was initially conducted using classical test statistics, more recently Rasch-based methods (see, e.g., Roznowski & Reith, 1999) have come to be the preferred statistical mode of analysis. A useful extension of DIF may be seen in 'bundling', that is grouping items into sets that share the same latent trait (e.g., Gierl et al., 2001). 'Bundling' in Rasch analysis (Linacre, 2012) is referred to as Differential Group Functioning (DGF), and in the case of test development purposes, DGF may be seen to be procedurally more informative than DIF (see Linacre, 2012). For ease of reference, however, given the general acceptance of the term "DIF", it is "DIF" that will be referred to in the current study to maintain consistency.

Depending on the type and level of test, it is not unlikely that some DIF will be found among certain background variables. While some studies investigating DIF have reported zero DIF (Chen & Henning, 1985), other studies have reported quite high incidences: Abbott (2004), for example, reports DIF of 62%. And even in studies where DIF has been found, a deeper exploration of DIF does not necessarily indicate any actual difference in performance between different DIF groups (Prieto & Nieto, 2014).

Data in the Current Study

DIF reported on in the current study was conducted on the objectively-marked Listening and Reading components of tests delivered by LanguageCert between 2018-2020. LanguageCert produce and administer a suite of exams – the International ESOL suite – which are aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The examination specifications reflect the six levels of the CEFR with regard to language attributes such as grammar, functions, vocabulary and discourse, function and how these relate to communication. Each exam has 52 items, of which 26 focus on reading and 26 on listening.

Since DIF optimally requires large sample sizes (Linacre, 2012), an exam with large sample sizes was identified for each CEFR level. Four background demographic variables have been used in the current study to explore DIF. Three of these – *gender*, *age*, *mother tongue* – are data supplied (optionally) by candidates upon registering for an exam. Given that many LanguageCert exams are conducted remotely through online proctoring (OLP), *test centre* is also taken as a variable. To make the analysis tractable, some recoding has been necessary, as laid out below.

- Gender: coded male / female.
- Age: recoded into decade of birth
- Mother tongue: only analysed where the sample size is greater than 10 incidences.
- Test centre:
  (1) analysed as is
  (2) recoded into either test taken face to face at a centre / test conducted via OLP

In the discussion and analysis below, DIF results are only presented for variables for which data exist; blank categories – i.e., where candidates did not report – have not been included. Further, only exam levels where DIF was observed are presented in the analysis. If a particular level does not appear in a table, that is because there was no DIF recorded for that level.

DIF investigations can operate at several levels. The aim of the present paper is to investigate initial DIF, i.e., the DIF of critical background variables. Such investigations provide evidence as to the overall quality and degree of bias in LanguageCert tests.

Results

In this section, Differential Group Functioning (DGF) using the computer program Winsteps (Linacre, 2010) is applied. Zwick (1999) provides an interpretation of significance (see also Linacre, 2010), where DIF strengths are graded into three categories, as in Table 1.

Table 1. DIF Strengths (after Zwick, 1999)

DIF Category	Strength	Logit size	Significance value
A	Negligible		
B	Slight to moderate	> 0.43 logits	$p < 0.05$
C	Moderate to large	> 0.64 logits	$p < 0.05$

In the discussion below, the focus will therefore be on Category C, moderate-to-large DIF. For brevity's sake, only overall summaries are presented for each test level.

Gender

Table 2 presents the results for gender. The reader's attention is drawn, as mentioned, to Category C – moderate-to-large DIF.

Table 2. DIF by Gender

Level		A	B	C	Total
A1	No.	20	3	1	24
	% within row	83.33 %	12.50 %	4.17 %	100.00 %
A2	No.	22	2	0	24
	% within row	91.67 %	8.33 %	0.00 %	100.00 %
B1	No.	11	3	1	15
	% within row	73.33 %	20.00 %	6.67 %	100.00 %
B2	No.	23	0	1	24
	% within row	95.83 %	0.00 %	4.17 %	100.00 %
C1	No.	22	1	1	24
	% within row	91.67 %	4.17 %	4.17 %	100.00 %
C2	No.	5	1	0	6
	% within row	83.33 %	16.67 %	0.00 %	100.00 %
Total	No.	103	10	4	117
	% within row	88.03 %	8.55 %	3.42 %	100.00 %

As can be seen, there are very few instances of DIF in Category C – 3.4% of the total.

Mother Tongue

LanguageCert has a list of over 100 mother tongues. The majority of these categories in the current dataset were either empty or had only one or two entries. Analysis has therefore only been conducted, as mentioned, where the sample size was greater than 10.

Table 3 below reports on the incidence of DIF totals for the 6 exam levels.

Table 3. DIF by Mother tongue

Level		A	B	C	Total
A1	No.	36	3	9	48
	% within row	75.00 %	6.25 %	18.75 %	100.00 %
A2	No.	67	12	9	88
	% within row	76.14 %	13.64 %	10.23 %	100.00 %
B1	No.	20	8	12	40
	% within row	50.00 %	20.00 %	30.00 %	100.00 %
B2	No.	72	13	19	104
	% within row	69.23 %	12.50 %	18.27 %	100.00 %
C1	No.	61	8	3	72
	% within row	84.72 %	11.11 %	4.17 %	100.00 %
C2	No.	8	0	0	8
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
Total	No.	264	44	52	360
	% within row	73.33 %	12.22 %	14.44 %	100.00 %

There is some incidence of DIF, with 14.4% of DIF reported for the C grade. In part this may be attributed to the wide scattering of different first languages and low sample sizes.

Decade of Birth

Year of birth may be seen to be an even more multifaceted variable than mother tongue. To this end, year of birth has recoded into decade of birth: 1960, 1970, 1980 etc. Table 4 presents the results.

Table 4. DIF by decade of birth

Level		A	B	C	Total
A1	No.	28	3	9	40
	% within row	70.00 %	7.50 %	22.50 %	100.00 %
A2	No.	42	7	7	56
	% within row	75.00 %	12.50 %	12.50 %	100.00 %
B1	No.	150	25	35	210
	% within row	71.43 %	11.90 %	16.67 %	100.00 %
B2	No.	43	3	2	48
	% within row	89.58 %	6.25 %	4.17 %	100.00 %
C1	No.	37	3	0	40
	% within row	92.50 %	7.50 %	0.00 %	100.00 %
C2	No.	10	0	0	10
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
Total	No.	310	41	53	404
	% within row	76.73 %	10.15 %	13.12 %	100.00 %

The incidence of DIF is 13.1%. From an examination of the data, there is no clear pattern of age or level.

Centre: Face to face vs. OLP

Over 200 centres around the world conduct tests in face-to-face mode. However, many of these conduct a very few tests.

Table 5. DIF by Centre

Level		A	B	C	Total
A1	No.	62	19	31	112
	% within row	55.36 %	16.96 %	27.68 %	100.00 %
A2	No.	99	18	35	152
	% within row	65.13 %	11.84 %	23.03 %	100.00 %
B1	No.	128	13	44	185
	% within row	69.19 %	7.03 %	23.78 %	100.00 %
B2	No.	16	0	0	16
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
C1	No.	16	0	0	16
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
C2	No.	4	0	0	4
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
Total	No.	325	50	110	485
	% within row	67.01 %	10.31 %	22.68 %	100.00 %

The incidence of C grade DIF is 22.7%. In part, this may again be attributed to the large number of centres, with some administering tests to a very small number of candidates.

It is difficult to comment objectively on DIF across centres since there are over 200 LanguageCert centres around the world. All centres are face-to-face institutions, with the exception of the LanguageCert centre which operates out of Athens, and which conducts exams remotely via OLP with candidates who are potentially of B2-C2 level. To shed some more light on the centre issue – and to the remote delivery of English language tests – a further focused analysis is now presented.

Centre: Face to face vs. OLP

LanguageCert is becoming a key player in delivering tests remotely through online proctoring (OLP). OLP is used to administer approximately 50% of LanguageCert's English language tests from the Athens centre. Against this backdrop and the multiplicity of centres, many of which have a very few candidates, *centre* has been recoded into OLP / face to face. Table 6 presents the DIF results for this analysis.

Table 6. Centre – OLP vs. face to face exam delivery

Mode	Level		A	B	C	Totals
OLP	B2	No.	8	0	0	8
		%	100%	0%	0%	100%
OLP	C1	No.	8	0	0	8
		%	100%	0%	0%	100%
OLP	C2	No.	2	0	0	2
		%	100%	0%	0%	100%
OLP	Total	No.	18	0	0	18
		%	100%	0%	0%	100%
face to face	B2	No.	8	0	0	8
		%	100%	0%	0%	100%
face to face	C1	No.	8	0	0	8
		%	100%	0%	0%	100%
face to face	C2	No.	2	0	0	2
		%	100%	0%	0%	100%
face to face	Total	No.	18	0	0	18
		%	100%	0%	0%	100%

From the aggregated analysis, no instances of C (nor of the less severe B) grade DIF were reported, either face to face at a physical centre or via OLP. Given the importance that LanguageCert attaches to its online proctoring operation, it is crucial that no bias be attached to this mode of delivery. The results in Table 6 above would appear to support this contention. ANOVA was used to investigate further and no significance bias was observed.

Reading or Listening

The IESOL examinations each comprise an equal number (26) of reading and listening items. From an investigation of DIF across both skills, it was concluded that there was no significant DIF in either reading or listening.

Conclusion

This study has investigated the incidence of DIF, or DGF (i.e., bundled DIF) across four of the background and location variables related to LanguageCert's suite of IESOL exams. The key focus of analysis has been on Zwick's moderate-to-large defining of DIF of 0.64 of a logit or greater. The overall preliminary finding based on the six exams analysed is that DIF is predominantly in Category A (negligible). A summary of the analysis of the four key variables explored is presented below.

- For *mother tongue*, a diverse category comprising over 100 languages, 15% moderate-to-large DIF was reported. Much of this may be attributable to the fact that many categories have only a very few entries.
- For *decade of birth* (recoded from year of birth), 13% moderate-to-large DIF was reported. Although decade of birth is a rather crude measure, it was used due to the small sample sizes. When larger sample sizes are available for analysis, it will be interesting to investigate in more depth.
- For *gender* – typically a key variable in the exploration of DIF – a very low incidence of 3% DIF was reported.
- Given LanguageCert's strong presence in remote delivery of tests, *centre* – comparing OLP vs non-OLP centres) – was also considered a variable of importance. Zero and moderate-to-large DIF was recorded for this variable.

In closing, it is worth comparing the incidence of DIF revealed in the current study with Ferne & Rupp's (2007) meta-analysis of DIF studies. The studies reported by Ferne & Rupp were essentially all tightly focused, that is, usually a single test with the focus on a single variable examining two clearly contrastive groups. A wide range of DIF across different studies was reported, as mentioned above.

The current study has involved the investigation of DIF over six ability levels (as per the CEFR), in the context of four background variables, two of which have 100 or more sub-categories. In this light, it may not be surprising that a degree of DIF was observed. However, while a degree of DIF was observed, it was encouraging to note that DIF related to gender and OLP versus centre delivery, was negligible.

In light of Prieto & Nieto's (2014) claims that DIF does not necessarily impact on overall candidate performance, our preliminary conclusion is that LanguageCert exams in this study are relatively reflecting the fact that they are carefully and professionally developed. Equally reassuring is the finding that there was minimal observable DIF in reading or listening components. Results generated from the six LanguageCert IESOL exams in this study may be seen as fair in relation to gender and delivery mode. Findings related to mother tongue and age are less compelling, but this is largely due to sample size.

References

Abbott, M. (2004). The identification and interpretation of group differences on the Canadian Language Benchmarks Assessment Reading Items. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Aryadoust, V., Goh, C.C.M., & Kim, L.O. (2011). An investigation of differential item functioning in the MELAB Listening Test. *Language Assessment Quarterly*, 8(4), 361-385.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155–163.

Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, 153(12), 1222–1226.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 113-148.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement Issues and Practice*, 20(2), 26-36.

Linacre, J. M. (2010). WINSTEPS, Version 3.69. Chicago, IL: Winsteps.com.

Linacre, J. M. (2012). A user's guide to WINSTEPS. Chicago, IL: Winsteps.com.

Prieto, G., & Nieto, E. (2014). Influence of DIF on differences in performance of Italian and Asian individuals on a reading comprehension test of Spanish as a foreign language. *Journal of Applied Measurement*, 15(2), 176-188.

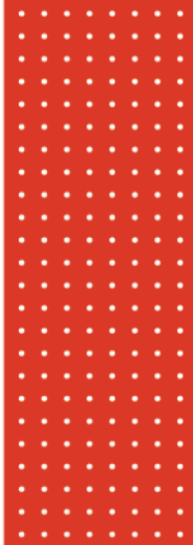
Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, 59, 248–269.

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, 9, 12–29.

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

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, 17, 323–340.

Zwick, R., Thayer, D. T., Lewis, C. 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36(1), 1-28.



Language Cert

languagecert.org