Language Cert

RECALIBRATING AND EXTENDING THE ANALYSIS OF THE LANGUAGECERT TEST OF ENGLISH

Nigel Pike

Yiannis Papargyris

Corina Dourda

Tony Lee

David Coniam

Abstract

This paper summarises a number of validation research projects carried out on the LanguageCert Test of English (LTE). Undertaken over a number of years, this work underpins the creation of the underlying LanguageCert Item Difficulty and Global Scales and aims to provide a single source of confirmatory evidence that the LTE system is a robust measurement tool in both linear and adaptive forms.

The paper extends and builds on analyses and calibration of the LTE system and, in particular, the adaptive test. This extensively-used test is drawn from the LTE item bank, which is also used to generate linear paper-based LTE tests. The particular adaptive test bank referenced in this paper consists of over 800 items and provides the basis for the studies reported below.

Keywords: LanguageCert Test of English (LTE), validation, item bank, adaptive test

Introduction

When effectively constructed and managed, item banks allow for the creation of test forms which are consistent and comparable both in terms of content and difficulty. This is relevant not only when creating tests intended to measure at a particular level (at CEFR level B1, for example) but also when developing tests which measure across multiple levels from A1 to C2. The master LTE item bank contains thousands of items, calibrated in terms of their difficulty on the LID scale which runs from 50-170. Candidate results are reported against the CEFR (the Common European Framework of Reference for Languages) levels, as well as the LanguageCert Global Scale which is aligned to the CEFR as laid out in Table 1 below. This scale is used with the full range of LanguageCert tests and allows for level comparison between tests, where appropriate and alignment of the tests to the CEFR for ease of reference.

Table 1: CEFR level, Global Scale results reporting and the LID scale

CEFR level	Global Scale score	LID scale range (item difficulty)	LID scale midpoint
A1	11-19	51-70	60
A2	20-39	71-90	80
B1	40-59	91-110	100
B2	60-74	111-130	120
C1	75-89	131-150	140
C2	90-100	151-170	160

The LID scale was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. Subsequent phases of measurement scale development for LTE build on the original LanguageCert Item Difficulty scale using Rasch analysis – in addition to expert judgement and CTS. The enhanced LID scale forms the empirical basis for the alignment of all current and future LanguageCert assessments to the same measurement scale that is itself aligned to the Common European Framework of Reference for Languages (CEFR).

All of the studies in this paper have the objective of establishing the robustness of the items used in the LTE tests, and the candidate results which emerge from the administration of the adaptive test (for an overview of the functioning of the adaptive test, see Pike and Coniam, 2021). The first section below describes the initial calibration studies; the following section outlines two simulation studies aimed at evidencing the stability of this adaptive test. The first simulation study explored potential future item bank stability via imputing and analysing a larger dataset; the second involved constructing tests from the item bank, administering those then-live tests to target sets of candidates and analysing the outcomes, i.e., candidate and item performance. Both studies indicate a robust item bank.

In addition to the perspectives of robustness and stability as judged by item and test quality, two studies report on candidates and their backgrounds. The first provides a picture of the composition and background demographics of candidates who have taken the LTE over the three-year period 2020-2023. The second study explores potential bias among candidates in terms of whether any of the eight item types was unfairly disadvantaging any subgroup of candidates.

Initial Calibration Studies

With a view to providing background to the analysis conducted, this section reports on four related studies.

Phase 1 of the analysis (Coniam et al., 2021a) took place in early 2021, and involved an analysis of four level-agnostic (i.e., which generated results from A1 to C2) paper-based (PB) tests comprising 364 items which had been administered to over 2,000 candidates in a number of countries. This study established a baseline measurement scale. Having calibrated the four tests onto a single scale using Rasch measurement, the embryonic scale was then aligned to the original LID scale. Rescaling the calibrated scale from standard logit values to a mid-point of 100 with a spacing factor of 20 resulted in a scale which was comparable to the original LID/CEFR level scale.

The calibrated Rasch scale produced from the four LTE paper-based tests which were seen to be well aligned to the LID scale then provided the baseline for further integration of LanguageCert products onto the common scale and validated the use of expert judgement and CTS in the original LID scale creation. All the items in the four paper-based tests are drawn from the overall LTE item bank and many of the items also feature in the adaptive test.

Phase 2 (Coniam et al., 2021b) involved an analysis of the adaptive test, which in mid 2021 consisted of over 800 items and 5,870 candidates. In the results, item and person reliabilities were both high. Rasch fit statistics – item and person infit and outfit mean squares – were well within acceptable ranges (i.e., 0.5 – 1.5), with the calibration statistics pointing to a test that could be viewed as sound.

It is worth noting that the calibration of the adaptive test – in terms of both item and candidate numbers – led to an improvement in the rigour of the LID scale with regard to percentile ranges and item distribution means. The scale mid-point (the 50th percentile) was 100 (99.92), closely matching the item distribution mean of 100.76. Following on, and everything else being equal, the mid-range ability group would be expected to occupy the major central region of the distribution while the higher and lower ability groups would be expected to occupy the upper and lower narrower range of ability. This indeed emerged to be the case: levels A1 and A2 fell below the 25th percentile, levels B1 and B2 between the 25th and 75th percentiles, and C1 and C2 in the top 25th percentile.

This positive picture notwithstanding, the sample size of 5,870 candidates was not considered to be sufficiently large to make definitive predictions about the robustness of the adaptive test. To this end, two approaches were seen as necessary. First, simulation studies (involving larger candidate sample sizes) would be conducted. Second, once the adaptive test had reached a comparatively large sample size (in the region of 50,000 candidates), the analyses in Phase 2 would be redone.

Confirming Item Bank Stability

With the purpose of examining the stability of the LTE adaptive test 1.0 from both statistical and operational perspectives, two simulation studies have been conducted with imputed large candidate sample sizes.

The first simulation study (Lee et al., 2022) was undertaken in late 2021, at which point, the adaptive test item bank comprising 827 calibrated items had been administered to over 13,000 candidates, each of whom had taken 58 items. In the study, performance in the 13,000-candidate live dataset was compared with a simulated much larger dataset generated using model-based imputation. Simulation regression lines showed a good match and Rasch fit statistics were also good: indicating that items comprising the adaptive test could be seen to be of high quality both in terms of content and statistical stability. Potential future stability was confirmed by results obtained from a Bayesian ANOVA.

The second simulation study (Coniam et al., 2022) built on the previous study, although with a different – real-world – focus, i.e., producing live tests from the LTE adaptive test, administering them to actual candidates and analysing the results. This process therefore involved submitting the adaptive test to a real-world test in that the quality of actual tests derived from the adaptive test was scrutinised. Three paper-based tests were compiled from the calibrated adaptive test and administered to target candidate groups. In the analysis of the three tests, good fit statistics emerged, with high correlations between each test – an indicator of robust joint calibration and further evidence as to the stability of the adaptive test. The second simulation study concluded with the claim that the items comprising the adaptive test were well set, and that the master LTE item bank (in its entirety, that is) was sufficiently robust to be used as a clearing house from which many different tests could be constructed. The caveat nonetheless remained that the analysis needed to be redone once a large candidate sample size – in the region of 50,000 – had been reached.

Confirming Fitness for Purpose

As of mid 2022, the adaptive test (comprising 827 items) had been administered to over 48,000 candidates. The studies described below are designed to confirm that the measurement characteristics remain stable with high volumes of candidates and that the item types used are fit for purpose.

The first recalibration study reported below (recalibrating the LTE adaptive test) builds on the research and analysis reported above, with two studies reported upon. This study updates the mid 2021 initial calibration study, which comprised 827 items and 5,870 candidates.

The second study extends the scope of the analysis – from analysing all (827) items in the adaptive test as a single entity – to a more fine-grained analysis, exploring the relative difficulty of the four different listening and four reading item types in the adaptive bank.

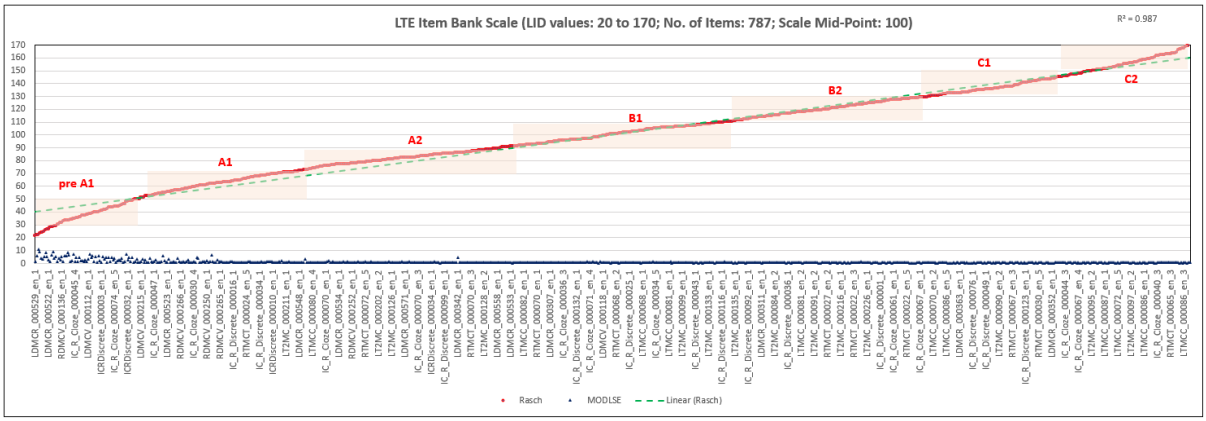
Adaptive Test (Re)calibration

As mentioned, as of mid 2022, over 48,000 candidates had each been administered 58 items via the adaptive test.

Following the methodology adopted in Lee et al. (2021), the 827 items were recalibrated using the midpoint of the scale (100, that is, B1) – in line with the previous calibration methodology. Of the 827 items, 21 were calibrated above 170 – the ceiling of the LID scale while 19 were calibrated below 20 – the bottom end of the LID scale. Those items were not included in the specification of the LTE scale presented below because the 21 items with values above 170 were too difficult for candidates while the 19 below 20 were too easy. Including such items in the final specification of the scale would have skewed distributions at both extreme ends. The final scale specification therefore currently has a total of 787 calibrated items.

Figure 1 below presents the picture the 787 items and their locations across the LID/CEFR levels.

Figure 1: Item distributions (N=787) across the item bank



The distribution of items, as presented in Figure 1, emerged at about 99% linear, especially in the A1 to C2 range. Such a distribution indicated a robust LTE scale with little distortion from the expected linear progression in an ability scale. Standard errors (SE) were minimal from A1 to C2. Even at pre-A1, where standard errors were highest, the largest SE was only 10 LID scale points, half a logit, a value commonly regarded as acceptable (Zwick, 1999).

Rasch Summary Statistics

Summary statistics for the dataset analysed via the Rasch measurement software Winsteps (Linacre, 2010) comprising 48,056 candidates and 827 items is presented in Table 2 below.

Table 2: LTE item bank summary statistics

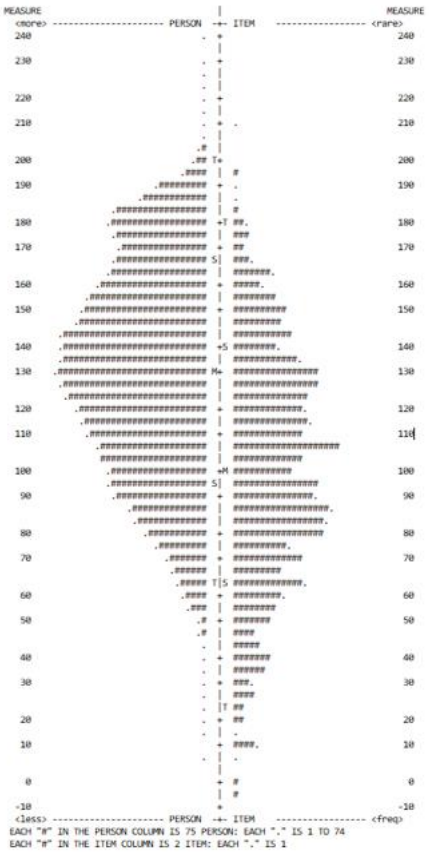
					INFIT		OUTFIT	
PERSON	48056	INPUT	48054	MEASURED				
	TOTAL		COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	36.3		56.7	131.31	6.52	1.00	.0	1.00	.0
P.SD	5.5		2.8	34.49	1.04	.12	.9	.39	.9
REAL RMSE	6.61	TRUE SD	33.85	SEPARATION	5.12	PERSON RELIABILITY			.96
ITEM	827	INPUT	827	MEASURED					
	TOTAL		COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	2109.7		3295.1	100.00	1.32	1.00	-.6	1.01	-.6
P.SD	1556.9		2122.6	39.00	1.50	.09	4.8	.19	5.0
REAL RMSE	1.99	TRUE SD	38.95	SEPARATION	19.56	ITEM RELIABILITY			1.00

Focusing on the right-hand, blue side of the table, item reliability was high at 1.00, as was person reliability at 0.96, the latter being the equivalent of classical test theory reliability (Anselmi et al., 2019). Person infit mean-square (1.00) and outfit mean-square (1.00) fit statistics were both within the acceptable range of 0.5 to 1.5, suggesting that the calibration of persons may be taken as acceptable. By the same token, item infit mean-square (1.0) and item outfit mean-square (1.00) fit statistics were also acceptable. Overall summary calibration statistics pointed, therefore, to a test that may be viewed as sound.

Person / Item Map

Person / item maps give a useful visual representation of candidate / item distributions. In Figure 2 below person/item maps are laid out such that the candidate spread (in LID scale points) appears to the left-hand side of the ruler while the item spread appears to the right-hand side of the ruler. More able candidates are located towards the upper left side of the map while less able candidates are located towards the lower left side of the map. Similarly, more difficult items are located towards the upper right side of the map while easier items are located towards the lower right side of the map.

Figure 2: LTE item bank person / item map



The item mean was set at 100; the candidate mean emerged at 131, the bottom of C1. The candidate mean was quite bell-shaped; the item mean showed a similar distribution, if slightly irregular in places.

Table 3 presents a distribution of item values by CEFR level.

Table 3: Item values at the CEFR levels

Item Type	Pre-A1	A1	A2	B1	B2	C1	C2
N	70	94	154	151	140	112	66
Mean	37.18	60.85	80.59	100.26	120.27	139.04	158.31
SD	7.93	5.49	5.63	5.98	5.87	5.85	5.53
Minimum	21.76	50.02	70.09	90.43	110.06	130.06	150.18
25th p'tile	31.61	56.67	76.51	95.23	115.31	133.97	153.56
50th p'tile	37.89	61.20	81.16	100.66	119.77	137.82	158.12
75th p'tile	43.91	64.85	85.51	105.90	125.59	143.68	162.68
Maximum	49.92	69.88	89.81	109.77	129.62	149.82	169.63

Minimum and maximum LID scale values for levels A1 to C2 emerged very close to the LID scale range values laid out in Table 1 above based on expert-judgement-assigned values. Test means were also very close to the midpoint of each level on the LID scale. This suggests that items are assessing at the desired levels.

A key statistic in the interpretation of Rasch results is that of data 'fit', which relates to how well obtained values match expected values (Bond et al., 2020). Broad criteria in assessing model fit are the *infit* and *outfit* mean square statistics (i.e., estimates of population variance, or standard error). *Infit* is generally seen as the 'big picture' in that it scrutinises the internal structure of an item. High infit values indicate rather scattered information within an item, providing a confused picture about the placement of the item. *Outfit* gives a picture of 'outliers' – responses from items which appear to be out of line with where an item would expect to be located.

For both infit and outfit, a perfect fit of 1.0 indicates that obtained values match expected values 100%. While acceptable ranges of tolerance for fit vary, acceptable ranges are generally taken as from 0.5 for the lower limit to 1.5 for the upper limit (Lunz & Stahl, 1990). 1.5 to 2.0 is considered just about acceptable, with figures beyond 2.0 unacceptable.

Table 4 presents the infit and outfit statistics for each CEFR level.

Table 4: CEFR level fit statistics

	A1		A2		B1		B2		C1		C2	
	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit
Valid	94	94	154	154	151	151	140	140	112	112	66	66
Mean	0.98	1.00	0.99	1.00	0.99	0.99	1.00	1.01	1.01	1.02	1.03	1.03
SD	0.06	0.20	0.09	0.21	0.07	0.14	0.09	0.18	0.10	0.16	0.14	0.17
Minimum	0.87	0.69	0.87	0.66	0.84	0.61	0.84	0.61	0.79	0.68	0.77	0.73
Maximum	1.20	1.92	1.37	2.16	1.29	1.43	1.41	1.79	1.32	1.51	1.38	1.47

Fit statistics were good at level mean values. At the extreme ends of the scale, at the A level for example, there was a degree of misfit. This may possibly be a result of small sample response size for approximately 100 of the 827 items in the dataset.

Item Type Analysis

There are four different listening and four reading item types in the adaptive test. Tables 5 and 6 below detail the constructs assessed in each item type, the (expert judge) assumptions regarding the relative demands of each item type, and the number of items currently in the adaptive test.

Table 5 first presents a breakdown of the Listening item types.

Table 5: Listening item types: Background detail

Item type	Constructs assessed	Relative demands	No. of items
LDMCR	Understanding spoken utterances and identifying the most appropriate response. (A1/A2 & C2) and interactions (B1-C1); awareness of functional language	A1-C2	123
LDMCV	Understanding key information in short spoken utterances; a focus on numbers, dates, spellings, prices etc	A1-A2, some B1/B2	41
LT2MC	Understanding short conversations; identifying opinion (sometimes unstated at C levels), standpoint, course of action, agreement/disagreement etc	B1-C2	128
LTMCC	Understanding longer monologues/dialogues; identifying fact, detail and chronology of events etc at A2-B1, opinion, cause-effect, speaker intention (sometimes unstated at higher levels) at B2-C2	A2-C2	100

Listening item types assess a range of constructs: some at a basic, essentially factual level (identifying numbers and dates etc), while others assess at a higher cognitive level (identifying opinion, agreement/disagreement etc.) Of the 393 listening items, the majority are broadly multi-level; a comparatively small number (40) focus on lower-level constructs, targeting CEFR A1/A2 Levels.

Table 6 provides a breakdown of the Reading item types.

Table 6: Reading item types: Background detail

Item type	Constructs assessed	Relative demands	No. of items
IC_R_Discrete	Understanding of vocabulary, collocation, phrasal verbs, idioms etc	A1-C2	142
IC_R_Cloze	Lexico-grammatical knowledge; vocab, linkers, phrasal verbs, collocation etc	A1-C2	170
RDMCV	Understanding the main idea of very short texts	A1-B1	38
RTMCT	Understanding longer texts ranging from detail and fact at lower levels (A2-B1) to complex argumentation, writer intention, summarising statements, unstated opinion etc at (B1) B2-C2	A2-C2	85

Reading item types also assess a range of constructs: some at a factual level (understanding of vocabulary), while others, as with Listening, assess at a higher cognitive level (understanding writer intention, unstated opinion etc.). As with the Listening items, the majority of the 434 Reading items are multi-level; only a small number are aimed at CEFR A1/A2 Levels, focusing on lower-level reading constructs.

Tables 7 and 8 below present item type difficulty from an analysis of the responses of the 48,000 candidates to whom the items have been administered. It should be recalled that, for purposes of analysis, the test midpoint is set at 100 (B1), with an SD of 20 (refer back to Table 1).

Table 7 first provides the analysis of the four Listening item types. The final row contains the expert-assigned target level. For the sake of readability, LID values have been rounded up to whole numbers.

Table 7: Listening item type values

	N	Mean	SD	Target level
LDMCR	123	122	33	A1-C2
LDMCV	41	65	12	A1-A2, some B1
LT2MC	128	129	27	B1-C2
LTMCC	100	137	29	A2-C2

Taking the mean as a reference point, the LTMCC items were seen to be the most demanding, with a mean of 137, or low-mid C1. While this item type assesses across levels, it also assesses certain higher level listening skills. LDMCV in contrast, being pitched at A1-A2, emerged with a mean of 65, or A1. As expected, the standard deviation for this task type is also by far the lowest as the range of levels tested is much smaller.

Table 8 presents the analysis of the Reading item types.

Table 8: Reading item type values

	N	Mean	SD	Target level
IC_R_Discrete	142	115	33	A1-C2
IC_R_Cloze	165	120	41	A1-C2
RDMCV	38	61	23	A1-B1
RTMCT	85	122	35	A2-C2

Regarding reading item types, RTMCT emerged as the most demanding item type, with a mean of 122, i.e., mid B2. This was closely followed by the IC_R_Cloze item type. The easiest type was RDMCV, at 61 (pre-A1). This task type (similar to LDMCV) had the lowest standard deviation as again the range of levels tested with this item type is narrower than the other item types.

Candidate Demographics Analysis

As a lead-in to the differential item functioning (DIF) analysis which is provided in the following section, an overview of the makeup of candidates is first presented. This overview, along with a summary of demographics, gives a picture of candidates who sat the LTE adaptive test over the three-year period mid 2020 to early 2023. The overview comprises four major categories: CEFR level obtained, country, gender and age. Second, crosstabulations by CEFR Level against country and gender are presented.

Overview of Major Categories

In terms of CEFR level candidature figures, there were few candidates at the CEFR A levels. This is to be expected as LTE is also available as a paper-based test, covering levels A1-B1, which is more appropriate for lower-level candidates. 65% of candidates were at B2 level and above. While there was some variation in the candidatures at the different CEFR levels over the past three years, the patterns of achievement at the different levels were broadly constant.

Regarding country of origin, while candidates from over 100 countries sat the LTE, many country candidatures were very small. Three countries – Poland, France and Greece – accounted for the majority of the candidature. With the exception of Greece, the largest candidatures were seen at B2 level.

In term of gender split, females accounted for 58% of the candidature. From A1-B2, there were more females than males. At C1, the genders were equal. It was only at C2 that more males were observed than females.

With regards age, the under 40s accounted for almost 70% of the candidature. For all age groups – apart from the 41-50 group – B2 was the level most commonly obtained.

Differential Item Functioning Analysis

This section extends the crosstabulation analysis presented with an investigation of Differential Item Functioning (DIF) into the three key variables. DIF analysis involves an exploration of whether any subgroup of candidates sitting a test is being unfairly disadvantaged. The exploration of potential bias among subgroup types typically involves investigating variables such as gender, first language, age etc. (Ferne & Rupp, 2007).

Rasch-based methods (Roznowski & Reith, 1999) have come to be the preferred statistical mode of analysis for DIF in terms of identifying latent traits. One extension of DIF is Differential Person Functioning (DPF), which involves the grouping of items into sets that share the same latent trait (e.g., Gierl et al., 2001). With over 800 items in the adaptive test, it was decided not to focus on the item level in this study. Rather, item groups are seen to be procedurally more informative and better indicators of both candidate performance and item precision than DIF (Linacre, 2012). DPF reports biases between candidates' actual responses against the estimated Rasch-calibrated item locations. Given the general acceptance of the term "DIF", however, it is "DIF" that is referred to in the current study.

The study follows the methodology described in Coniam & Lee (2021), where bias was investigated in LanguageCert IESOL Listening and Reading tests. In the current study, analysis has been conducted using the computer program Winsteps (Linacre, 2010). Since 100 is the mid-point of the LanguageCert Item Difficulty scale (see Table 1 above), Rasch logit values are rescaled to a mean of 100 and a standard deviation (SD) of 20 (Coniam et al., 2021b). As mentioned, DPF involves bundling items together; the analysis is therefore conducted on the basis of the four Reading and four Listening item types.

Three key statistics are reported in the analysis below. These are laid out in Table 9.

Table 9: Key statistics reported in DIF analyses

Statistic	**Gloss**	**Comment**
N	Number of responses analysed	
Item Facility (IF)	Percentage of correct responses	0.50 is taken as the IF threshold: an indicator that candidate correct responses were not successful merely by chance.
DIF Size	Difference between actual and Rasch calibrated locations	Positive values indicate that candidate responses were higher than calibrated values, and vice versa.

In analytic terms, DIF strengths may be graded into three categories: A, B and C (Zwick, 1999). 'A' indicates negligible DIF while 'C' is the most demanding category, indicating moderate-to-large DIF (greater than 0.64 logits). In the study, the threshold of 10 LID scale points, or half a logit, is taken as the limit for indicating possible biased responses.

DIF Analyses

A detailed summary of the DIF analyses is presented below. In the analysis of item type against gender, country and age, no DIF greater than 10 LID scale points (half a logit) on any of the three variables analysed was reported. No Category C, moderate-to-large DIF (Zwick, 1999), was observed.

In the tables below, DIF size biases above (or close to) 10 LID scale points are highlighted in red.

Table 10: DIF by gender

Analysis					Commentary
Gender	**Item type**	**N**	**IF**	**DIF size**	
F	LDMCR	361548	0.74	-0.82	All Item Facilities (IF) are above 0.5, so it may be taken that candidate correct responses were not merely chance guesses. All DIF sizes are small, indicating that there would appear to be no bias regarding gender in the LTE data.
F	LDMCV	16324	0.78	-0.4	
F	LT2MC	263231	0.62	-0.7	
F	LTMCC	136009	0.59	0	
F	IC_R_Discrete	362464	0.55	0.72	
F	IC_R_Cloze	277912	0.62	0	
F	RDMCV	25304	0.57	0	
F	RTMCT	138801	0.65	0	
M	LDMCR	252237	0.75	1.11	
M	LDMCV	11601	0.78	0	
M	LT2MC	183416	0.64	1	
M	LTMCC	95039	0.61	0.42	
M	IC_R_Discrete	248794	0.59	-1.17	
M	IC_R_Cloze	193919	0.65	-0.68	
M	RDMCV	16454	0.58	-0.76	
M	RTMCT	96882	0.66	0	
n/a	LDMCR	10218	0.71	0.51	(n/a = not available)
n/a	LDMCV	680	0.79	0.77	
n/a	LT2MC	7227	0.6	0	
n/a	LTMCC	3766	0.57	-1.02	
n/a	IC_R_Discrete	10465	0.55	0.57	
n/a	IC_R_Cloze	7860	0.6	0	
n/a	RDMCV	968	0.58	-0.6	
n/a	RTMCT	3921	0.64	-2.04	

Table 11: DIF by Country

Analysis					Commentary
Country	**Item type**	**N**	**IF**	**DIF size**	All Item Facilities (IF) (with one 0.49 in the Poland data) are again above 0.5, so it may be taken that candidate correct responses were not by chance. There does not seem to be a country bias in the LTE adaptive test data.
Germany	LDMCR	5842	0.75	0	
Germany	LDMCV	102	0.84	-8.99	
Germany	LT2MC	4420	0.64	-1.31	
Germany	LTMCC	2167	0.61	-2.17	
Germany	IC_R_Discrete	6167	0.54	1.88	
Germany	IC_R_Cloze	4500	0.62	0.63	
Germany	RDMCV	237	0.64	-2.68	
Germany	RTMCT	2239	0.67	-2.65	
Italy	LDMCR	9641	0.70	0.00	
Italy	LDMCV	716	0.78	1.86	
Italy	LT2MC	6751	0.59	0.00	
Italy	LTMCC	3592	0.58	-2.47	
Italy	IC_R_Discrete	9856	0.54	1.19	
Italy	IC_R_Cloze	7416	0.59	0.00	
Italy	RDMCV	1005	0.57	0.00	
Italy	RTMCT	3706	0.64	-3.18	
Poland	LDMCR	70220	0.73	-2.36	
Poland	LDMCV	3125	0.78	-1.54	
Poland	LT2MC	51226	0.64	-4.21	
Poland	LTMCC	26628	0.61	-4.62	
Poland	IC_R_Discrete	73241	0.49	5.05	
Poland	IC_R_Cloze	54009	0.58	2.51	
Poland	RDMCV	5000	0.56	0.73	
Poland	RTMCT	27002	0.66	-4.31	
France	LDMCR	178973	0.68	0	
France	LDMCV	12115	0.76	0	
France	LT2MC	126595	0.59	-0.89	
France	LTMCC	66882	0.55	-3.06	
France	IC_R_Discrete	182742	0.54	0.58	
France	IC_R_Cloze	137463	0.58	1.19	
France	RDMCV	18466	0.56	0	
France	RTMCT	68621	0.64	-0.51	
Greece	LDMCR	341757	0.77	0.00	
Greece	LDMCV	11831	0.79	0.00	
Greece	LT2MC	252034	0.65	1.39	
Greece	LTMCC	128948	0.62	2.94	
Greece	IC_R_Cloze	262831	0.67	-1.21	
Greece	RDMCV	16976	0.59	-0.60	
Greece	RTMCT	131298	0.66	1.33	
Greece	IC_R_Discrete	331518	0.60	-1.65	
Other	LDMCR	17570	0.74	1.37	
Other	LDMCV	716	0.80	-1.16	
Other	LT2MC	12848	0.63	0.00	
Other	LTMCC	6597	0.59	-1.13	
Other	IC_R_Discrete	18199	0.55	0.71	
Other	IC_R_Cloze	13472	0.64	-1.98	
Other	RDMCV	1042	0.58	0.99	
Other	RTMCT	6738	0.67	-0.76	

Table 12: DIF by Age

Analysis					Commentary
Age	**Item type**	**N**	**IF**	**DIF size**	There would not appear to be any bias as regarding age.
under 31	LDMCR	203828	0.70	0.00	
under 31	LDMCV	11829	0.77	-0.75	
under 31	LT2MC	146001	0.61	-1.15	
under 31	LTMCC	76783	0.58	-2.97	
under 31	IC_R_Discrete	207956	0.53	1.38	
under 31	IC_R_Cloze	156660	0.59	1.08	
under 31	RDMCV	18124	0.56	0.00	
under 31	RTMCT	78260	0.64	-0.76	
31-40	LDMCR	216352	0.75	0.00	
31-40	LDMCV	9729	0.78	0.70	
31-40	LT2MC	157526	0.64	0.00	
31-40	LTMCC	81735	0.61	0.63	
31-40	IC_R_Discrete	212635	0.57	0.00	
31-40	IC_R_Cloze	166339	0.64	0.00	
31-40	RDMCV	13959	0.57	0.00	
31-40	RTMCT	83092	0.66	0.00	
41-50	LDMCR	117139	0.77	0.00	
41-50	LDMCV	3864	0.80	-1.46	
41-50	LT2MC	86621	0.65	0.49	
41-50	LTMCC	44124	0.62	1.74	
41-50	IC_R_Discrete	114675	0.58	-0.92	
41-50	IC_R_Cloze	90066	0.66	-0.43	
41-50	RDMCV	5791	0.59	-0.59	
41-50	RTMCT	44984	0.67	0.00	
51-60	LDMCR	62154	0.76	0.00	
51-60	LDMCV	1993	0.79	0.00	
51-60	LT2MC	45968	0.64	0.75	
51-60	LTMCC	23204	0.60	2.82	
51-60	IC_R_Discrete	62247	0.58	-0.97	
51-60	IC_R_Cloze	47790	0.66	-1.02	
51-60	RDMCV	3076	0.59	0.00	
51-60	RTMCT	23857	0.66	0.60	
over 60	LDMCR	23701	0.75	1.49	
over 60	LDMCV	1098	0.78	2.47	
over 60	LT2MC	17203	0.62	2.87	
over 60	LTMCC	8686	0.59	4.94	
over 60	IC_R_Discrete	23444	0.62	-2.87	
over 60	IC_R_Cloze	18198	0.68	-3.16	
over 60	RDMCV	1644	0.61	-1.07	
over 60	RTMCT	9099	0.65	1.42	
n/a	LDMCR	829	0.67	2.22	
n/a	LDMCV	92	0.75	**9.89**	(n/a = not available)
n/a	LT2MC	555	0.62	-2.99	
n/a	LTMCC	282	0.52	0.48	
n/a	IC_R_Discrete	766	0.58	-2.19	
n/a	IC_R_Cloze	638	0.59	1.12	
n/a	RDMCV	132	0.60	-3.21	
n/a	RTMCT	312	0.63	3.10	

Finally, to explore how well current results might hold in the future, a Bayesian equivalence t-test was run against the adaptive test and DIF scores. The results are provided in Table 13.

Table 13: Bayesian equivalence t-test run on LTE adaptive test scores and DIF values

					95% Credible Interval	
	N	**Mean**	**SD**	**SE**	**Lower**	**Upper**
LTE scores	319486	127.28	35.33	0.06	127.16	127.40
DIF values	319486	127.96	40.62	0.07	127.82	128.10

In Table 13 above, the means of both sets of values together with credible interval values in the LID scale range of 127 (see Table 1) are located in the middle of the B2 range. The LTE scores and DIF values scores may therefore be taken as equivalent within their respective credible intervals.

The conclusion that may be drawn from the DIF study is that LanguageCert tests are as bias free as one would wish against a backdrop of tests that are carefully and professionally developed. There was no predominance of DIF on either Reading or Listening item types against country, gender or age. Results generated from the LTE adaptive test may be therefore considered fair in the context of candidate background and language skill.

Conclusion

This paper has outlined background studies which have contributed from different perspectives to the calibrating of the LanguageCert LTE via the LID scale; and to how the LTE functions operationally in terms of candidate demographics and possible item bias. The LID scale is a comprehensive scale, linked to an item bank which provides both anchoring from individual tests with different frames of reference (Humphry, 2006) and individual item-based adaptive tests. Against this backdrop, the LanguageCert scale should be viewed as a hybrid scale – in that it provides the foundation for the development and creation of both standalone and adaptive tests.

The engine facilitating the construction of LanguageCert tests involves a complex item banking system containing large amounts of test material. This test material covers a wide range of content and construct characteristics which has been calibrated on the basis of Rasch difficulty estimates and fit statistics, and classical test statistics analysis.

When effectively constructed and managed, item banks allow for the creation of test forms which are consistent and comparable both in terms of content and difficulty. This is relevant not only when creating tests intended to measure at a particular level (at CEFR level B1, for example) but also when developing tests which measure across multiple levels from A1 to C2.

The current paper has outlined a number of related background studies, with two simulation studies conducted to ascertain item bank robustness. The first simulation study explored potential future item bank stability via imputing and analysing a larger dataset; the second simulation study involved a real-world test in terms of constructing tests from the item bank, administering those then-live tests to target sets of candidates and analysing the outcomes, i.e., candidate and item performance. Both studies contributed to a picture of a robust item bank.

In addition to the perspective of robustness as judged by item and test quality, two studies have reported on candidates and their backgrounds. The first provided a picture of the composition and background demographics of candidates who have taken the LTE over the three-year period 2020-2023. The second study explored potential bias among candidates in terms of whether any of the eight item types was unfairly disadvantaging any subgroup of candidates. The Differential Item Functioning investigation into the three key variables of country, gender and age reported no major item bias.

References

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, 10, 2714.

Bond, T., Yan, Z., & Heene, M. (2020). Applying the Rasch model: Fundamental measurement in the human sciences. New York: Routledge.

Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021a). Validating the LanguageCert Test of English scale: The paper-based tests. London, UK: LanguageCert.

Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021b). Validating the LanguageCert Test of English scale: The adaptive test. London, UK: LanguageCert.

Coniam, D., & Lee, T. (2021). Potential bias in LanguageCert IESOL items: A Differential Item Functioning analysis. London, UK: LanguageCert.

Coniam, D., Lee, T., & Milanovic, M. (2022). Exploring Item bank stability in the creation of multiple test forms. London, UK: LanguageCert.

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.

Humphry, S. (2006). The impact of differential discrimination on vertical equating. ARC report. Western Australia: Department of Education & Training.

Lee, T., Coniam, D., & Milanovic, M. (2022). Exploring item bank stability through live and simulated datasets. *Journal of Language Testing & Assessment*, 5, 13-21.

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.

Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. Eval*uation and the Health Profession*, 13, 425-444.

Pike, N. & Coniam, D. (2021). Adaptive testing and the LanguageCert Test of English adaptive test. *ELTNEWS*, 367.

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.

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