

Language Cert

The LanguageCert Global Scale



Michael Milanovic
Nigel Pike
Yiannis Papargyris
Tony Lee
and
David Coniam



Introduction

The LanguageCert system is based on a measurement scale that is aligned to the Common European Framework of Reference (CEFR) and can, in turn, be aligned to other scales as required. The scale has been in development for some years (since 2017) and as data is gathered from the range of LanguageCert assessments, the scale is subjected to on-going validation. As the requirements of users have become better defined, the nature of the underlying measurement scale has also developed and will progressively embrace the full range of LanguageCert exams.

This paper reports on the development of this measurement scale through a number of phases, culminating in what is now referred to as the *LanguageCert Global Scale*. We first provide background to the original LanguageCert scale – the LanguageCert Item Difficulty (LID) scale. We describe its development, implementation and calibration. Discussion then moves to the nature and purpose of the Global Scale and on to the transition from the LID to the Global Scale in terms of calibration and alignment.

Background to LID Scale

The initial LanguageCert Item Difficulty (LID) scale was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. LID scale difficulty values range from CEFR Pre-A1 through to high C2 level. The scale ranges and midpoints are presented in Table 1 below.

Table 1: LID scale

CEFR level	LID scale range	Midpoint
A1	51-70	60
A2	71-90	80
B1	91-110	100
B2	111-130	120
C1	131-150	140
C2	151-170	160

As mentioned, the LID scale was developed using both expert judgement and item analysis such that 20 points separated each CEFR level. In 2017, eight expert consultants, each of whom had over 20 years writing, editing and vetting test materials to measure directly against the CEFR, completed a standards-setting exercise which generated anchor material to enhance and validate the scale. These anchor items then underwent trials and live tests, with all other items in the LanguageCert item banks measured against them, thereby giving each item in these tests a difficulty value on the LID scale. An in-depth analysis was conducted on all anchor items at this stage and a small number were eliminated from further use as anchors, as they were not measuring as predicted. In the following sections, we summarise five studies that describe – against the backdrop of the LTE adaptive item bank – the validation of the LID scale.

Study 1: Initial Calibration of Paper-based Tests (2020)

One of the LanguageCert item banks is devoted to the LanguageCert Test of English (LTE). This test provided a very useful set of data in that it offers both linear and adaptive tests, measuring on the same scale. The bank used in these studies in 2020 contained, at the time, over 1000 items and was used to generate an adaptive test and linear tests.

To validate the expert judgements used to generate the original LID scale, a calibration exercise involving Rasch measurement was undertaken in 2020, with the focus on LTE. This version of the LTE is an English 'for work' exam intended for people over 18 in or about to enter the workplace, as well as those in higher or further education. It has been accredited by the UK's Office of Qualifications and Examinations Regulation (Ofqual). The LTE is available in three versions described in Table 2 below.

Table 2: LanguageCert Test of English versions

Test version	CEFR levels aimed at
(1) paper-based (PB) test measuring from A1-B1	beginner to intermediate CEFR levels
(2) PB test measuring from A1-C2	candidates at all CEFR levels
(3) adaptive test measuring from A1-C2	candidates at all CEFR levels

All three versions of the LTE are produced from the same LTE item bank. At the time of analysis (2020), the LTE item bank that was to be analysed contained around 1,600 items. Currently, it contains over 3,500 items and continues to grow. From this item bank, both paper-based and adaptive tests were produced, utilising in total approximately 1,600 items (827 in the adaptive test and more than 1,000 in the PB tests) with many common items between the CAT and the PB tests, and between different versions of the PB tests for cross-calibration purposes.

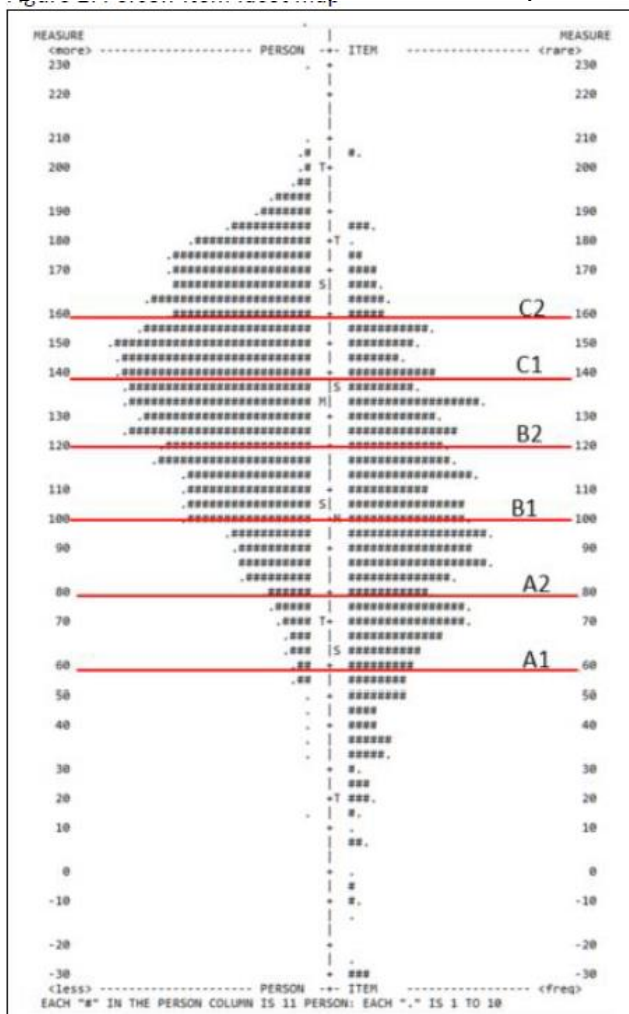
The first study explored four paper-based (PB) tests with a view to establishing an initial set of anchor items, by which the entire item bank might be subsequently calibrated. The initial sample comprised a total of 282 discrete items in the four-test database which had been administered to 2,112 candidates. The fit of the items to the Rasch model was good and reliability was high. With all four tests calibrated to a single scale, the calibrated scale was rescaled to a mid-point of 100 with a spacing factor of 20 in order to align the calibrated Rasch scale and the original LID scale. The rescaling of the Rasch scale in this manner produced a comparable alignment between the two scales although some differences were detected at the A level which required further exploration.

Study 2: Calibration of Adaptive Test Item Bank (2021)

The initial calibrated scale that emerged from the set of paper-based tests demonstrated that the paper-based tests were robust and consistent with the data. This provided a basis for the further validation of the LID scale through data generated by the adaptive test and was the second major calibration study.

The calibration conducted in 2021, based on the LTE item bank incorporated the 827 items in the LTE adaptive test which had been administered to 5,800 candidates (with each candidate having taken approximately 60 items). The dataset incorporated the 282 calibrated items from the paper-based tests in Study 1. These items formed anchors in the Rasch measurement calibration. The Rasch person-item map in Figure 1 below shows the fit of candidates (to the left-hand side of the map) and items (to the right).

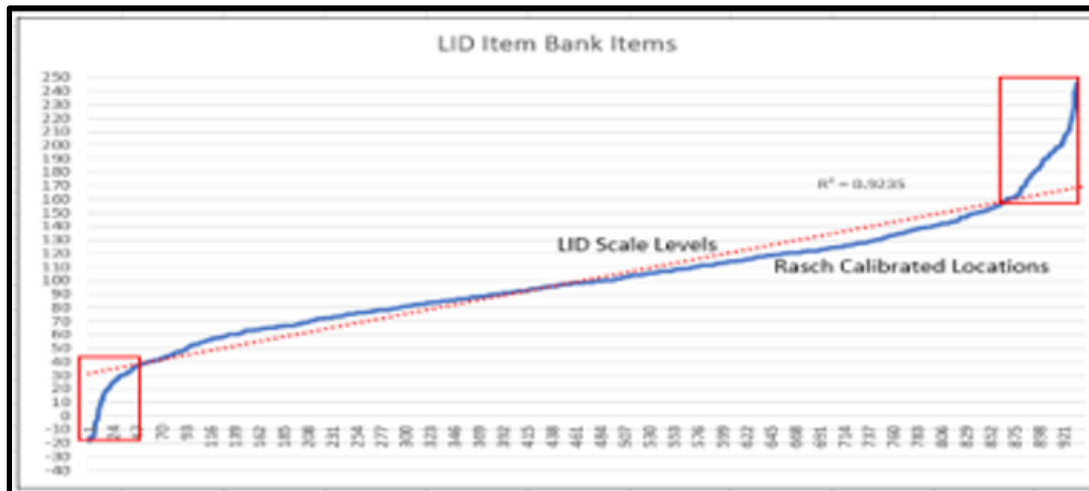
Figure 1: Person-Item map from LTE adaptive item bank calibration



As Figure 1 illustrates, person and item distributions extended approximately 120 points, or six logits – the rule-of-thumb operational range (Bond et al., 2020). Candidates generally matched with items. The person distribution is dependent upon the nature of the test population, and a considerable number of high ability candidates were known to be in the sample.

Concerning the items, there was nonetheless generally a good match between Rasch calibrated locations and LID scale expert-defined levels. Figure 2 illustrates.

Figure 2: Calibrated locations and expert-defined level fit in adaptive bank test items



While the R^2 figure was good at 0.9, there was a degree of misfit at the bottom (the 'lag' phase of the Sigmoid or 'S' curve [Handy, 1995]) and top (the 'steady state') ends of the scale, which should ideally be flat with small gradients to indicate slower rates of item difficulty increase. As can be seen in Figure 2, there are sudden rises and falls in item difficulty levels at the extreme ends of the scale, with items being either too easy (at the bottom end) or too difficult (at the top end). It was felt that the sharp downturn at the bottom end of the distribution was due for the most part to the fact that there were very few A1 and pre-A1 candidates in the dataset. The sharp upturn at the top of the distribution may have related to the inclusion of a small number of very difficult items.

However, the conclusion drawn from Study 2 was that the LID scale could be considered to be a comprehensive and robust scale.

Studies 3 and 4: Simulations (2022)

The next stage in the validation process was to consider the stability of the bank. With a coherent LID LTE scale developed, and when the adaptive test cohort surpassed 10,000 candidates, two linked studies exploring the stability of the 827 items in the adaptive test were conducted. Study 3 explored item bank stability through a simulated 'full' dataset generated through model-based imputation (i.e., whereby the parameter values of persons, items and thresholds from the current analysis were used to generate simulated data according to the probabilistic distributions defined by the Rasch model and generating Rasch parameters). Results pointed to item bank stability, indicating that items making up the adaptive item bank were of high quality both in terms of content and statistical stability. Potential future stability was confirmed by results obtained from a Bayesian ANOVA.

A linked follow-up study (Study 4) involved submitting the items to a 'real-world' test by which three (paper-based) tests were compiled from the calibrated items in the adaptive test and was administered to a sample of test takers. In the analysis of the three tests, good fit statistics emerged, with high correlations between the tests – an indicator of robust joint calibration and further evidence as to the stability of the item bank.

Study 5: Finalising the Calibration (2022)

As of mid 2022, the LTE adaptive test used in these studies comprised 827 items and had been administered to over 48,000 candidates. A recalibration was then performed. Figure 3 below summarises the recalibration results.

Figure 3: Summary of Rasch analysis

PERSON	48722	INPUT	48054	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE		IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	34.9	56.7	131.80	6.68		1.00	.0	1.01	-.0
P_SD	5.8	2.8	35.14	1.09		.12	.9	.42	-.9
REAL RMSE	6.77	TRUE SD	34.48	SEPARATION	5.10	PERSON RELIABILITY	.96		

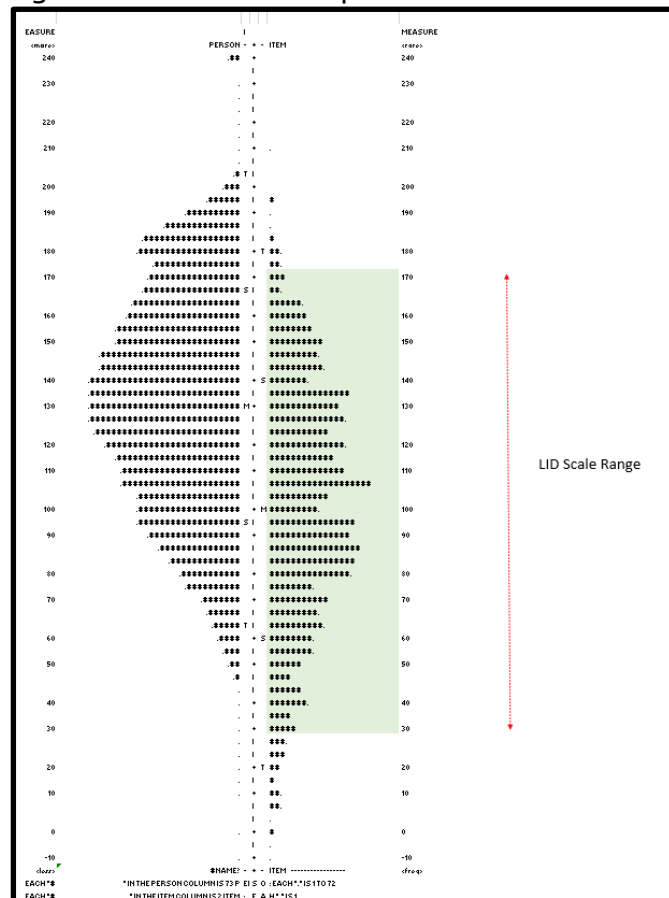
ITEM	923	INPUT	827	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE		IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	2027.8	3295.1	110.12	3.17		1.00	-.6	1.01	-.6
P_SD	1614.0	2122.6	59.04	7.96		.09	4.9	.20	5.1
REAL RMSE	8.57	TRUE SD	58.41	SEPARATION	6.81	ITEM RELIABILITY	.98		

Measurement error (RMSE) was 8.57 (less than half a scale level against the 20-point LID scale); the separation index (an index pointing to construct validity) of 6.81 was well above the customary decision level of 2.0 for good separation, indicating clearly distinguishable item locations with little chance of overlaps due to measurement error. Reliability was very high at 0.98.

To finalise the calibration, additional external calibration was conducted on A1 and C2 level items, with the results subsequently incorporated into the overall LID scale.

The item-person map incorporating all levels is presented in Figure 4 below.

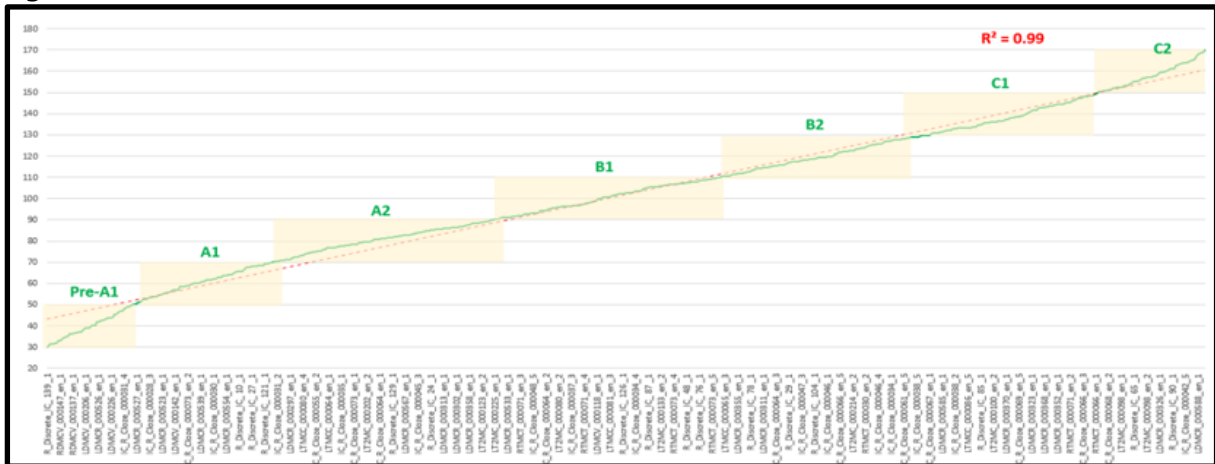
Figure 4: Person-Item map



As can be seen, LID scale item difficulties range from 30 (Pre-A1) to 170 (C2).

The finalised LID scale after calibration is presented in Figure 5.

Figure 5: Finalised LID scale calibration



As can be seen, there is a generally linear gradation from pre-A1 up to C2.

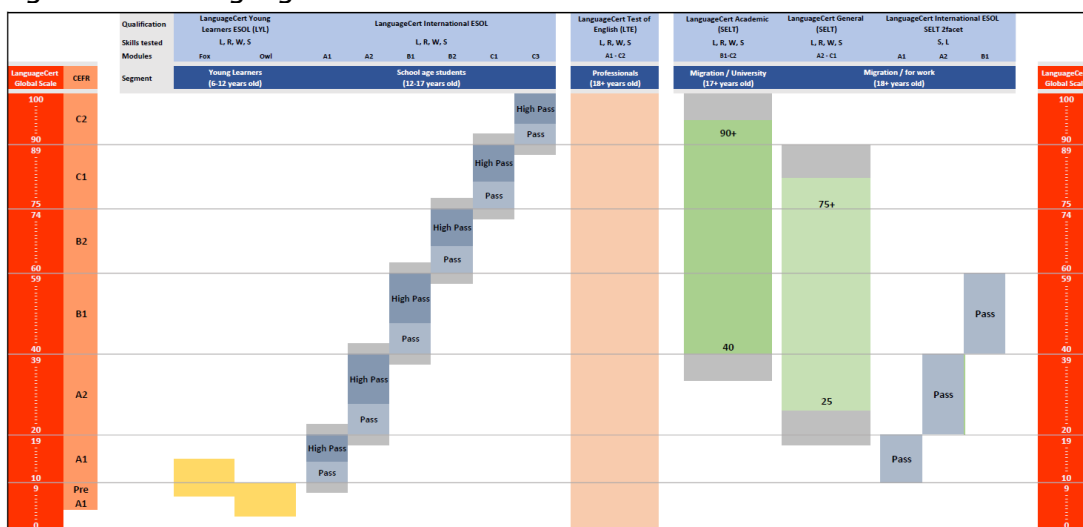
The results above provided confidence that the LID is measuring as has been claimed, with the 800+ LTE items in the adaptive test version forming the bedrock, the base, against which future items and tests may be calibrated.

The underpinning of the LID scale now allows for the transition to the Global Scale to be outlined, to which the discussion now turns.

Transitioning from the LID to the Global Scale

The LID scale was intended as an internal item difficulty scale in the item bank system. Feedback from external users of the LID scale suggested that the effective scale range of 50-170 was not sufficiently intuitive. Therefore, in order to make the scale easier to work with, following consultation, it was decided to recalibrate the scale to 0-100 and use this as a basis for mapping all LanguageCert tests. It was also felt that 'LID' was not very transparent as a name. As a consequence the scale was renamed the Global Scale. It links directly to the LID scale and thereby the CEFR levels. Performance on LanguageCert tests can then be mapped to other English language testing organisations' examinations such as IELTS and Cambridge Advanced. Figure 6 illustrates an initial representation of the Global Scale and how it reports against CEFR levels.

Figure 6: The LanguageCert Global Scale



The figure above illustrates how the LanguageCert System reports scores on the LanguageCert Global Scale of 0-100 and applies across all the tests in the LanguageCert System. The Global Scale provides candidates, employers, education institutions and government agencies an easy-to-understand results system. The Global Scale defines specific levels of attainment needed to fulfil certain requirements. For example, entrance into a university or for migration and employment purposes. The levels of attainment can relate to overall performance in an examination, performance by skill (e.g., speaking), or both these parameters.

Mapping the LID Scale to the Global Scale

Figure 6 presented a visualisation of the Global Scale in relation to the full range of assessments offered by the LanguageCert system. Before full implementation, however, it must be demonstrated that the LID scale, to which all LanguageCert tests have been aligned, maps cleanly and clearly to the Global Scale. The discussion below outlines how this issue has been addressed.

Two methods were considered regarding mapping the two scales to each other. The first method was to simply divide the active 120-point LID scale into 100 with 1.2 LID points per Global Scale level point. While this method might appear intuitive, the realignment of a 120-point to a 100-point level would be mathematically fraught in terms of actual administration. More importantly, such a realignment would result in an ordinal scale which progresses in integer steps, omitting in-between step differences, and hence possibly obscuring between-level differences, potentially misrepresenting scores. An interval scale, in contrast, is continuous and permits in-between step differences.

Having therefore discounted the simplistic calculation of 120 to 100 by 1.2 scale points, the methodology adopted was to calibrate the LTE item bank such that a scale mid-point and a logit value yielded a scale with 100 as total. The aim is therefore to shift the scale to a lower mid-point and with a narrower logit range in order to transition to the Global Scale.

After several iterations, the scale mid-point of 50.5 and a logit of 15 were found to yield a good approximation, with a Pearson correlation of 1.0 between the LID and Global scales. The Global Scale (GS) that emerged is presented in Table 3 below.

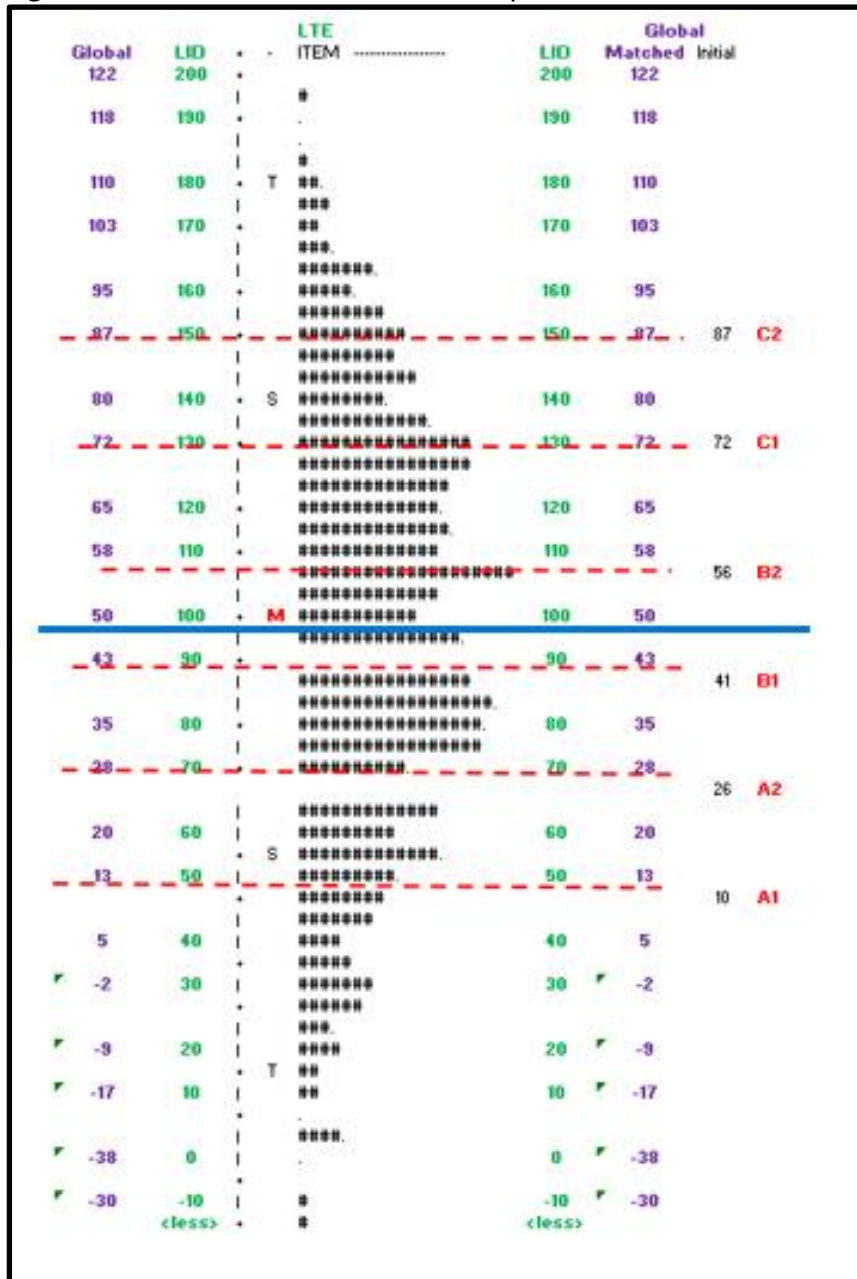
Table 3: LID and Global Scale fits

CEFR	LID Scale upper and lower ranges	Global Scale upper and lower ranges	GS cut score point	GS point range
C2	170.08	100		
C2	150.25	87.03	87	
C1	149.82	86.33		
C1	130.67	72.13	72	15
B2	129.93	71.40		
B2	110.06	56.36	56	16
B1	109.75	55.57		
B1	90.05	41.63	41	15
A2	89.79	40.83		
A2	70.01	26.05	26	15
A1	69.90	25.68		
A1	50.02	10.62	10	16

As can be seen, the Global Scale level widths are not uniformly equal. A1 begins at 10; there are 15 or 16 points up until C1, and then C2 comes in above 87.

With the new midpoint and spacing factor, the items in the LTE item bank needed to be recalibrated. The visual mapping of the two scales on the item map is presented in Figure 7 below.

Figure 7: LID and Global Scale item map



Future LanguageCert Assessment Products

For the recently introduced LanguageCert Academic and General tests (see Jones, 2023), results are reported against the CEFR levels and on the LanguageCert Global Scale. The Global Scale score (which is provided by language skill and overall result) gives finer gradations of performance within the CEFR levels but is also a standalone measure that can be aligned with any relevant external scale.

LanguageCert Academic and General Tests

The Listening and Reading tests in the LCA and LCG series of tests were calibrated with the calibrated Rasch results of the LID item bank as the calibration standard. This was achieved by linking the LCA and LCG tests via common items in the LTE item bank. In order not to dislocate established calibrated Rasch values in the LID scale, all items in the LID scale were anchored before calibration with the LCA and LCG tests appended to the dataset. Since the initial LID scale values have now been matched with the 100-point Global Scale, the combined LID LCA / LCG data were calibrated using the 100-point Global Scale as the reference scale. It should be noted that the process described will be the general process adopted as future LanguageCert tests are matched to the Global Scale.

Figure 8 below presents the calibrated LCA L&R, LCG L&R tests against the (re-calibrated) LTE item bank.

Figure 8: Global Scale showing re-calibrated LTE item bank and LCA L&R, LCG L&R tests

Global	LTE	LCA Lt	LCA Rd	LCG Lt	LCG Rd	Global
+	ITEM	+	ITEM	+	ITEM	
122	+	+	+	+	+	122
118	+	+	+	+	+	118
110	T	X	+	X	+	110
103	+	+	T XX	+	+	103
95	+	T	X	+	+	95
87	+	X	S	X	X	87
80	S	X	XX	X	S X	80
72	+	X	XXXXX	X	XXX	72
65	+	X	XX	X	XX	65
58	+	XX	XX	XX	M	58
50	M	XX	XX	XX	XX	50
43	+	X	+	M X	X	43
35	+	XX	+	XXX	+	35
28	+	X	X	X	X	28
20	+	X	+	X	T	20
13	+	S X	+	S	XX	13
5	+	X	+	X	+	5
-2	+	X	+	X	+	-2
-9	+	XX	+	XX	+	-9
-17	T	T	+	T	+	-17
-38	+	+	+	+	+	-38
-30	+	+	+	+	+	-30

S - ITEM
 E O : EACH ; : IS 1TD 74
 A H " : IS 1

Estimating Reliability

A widely adopted approach to derive overall scores for language tests comprising two or more of the four skills involves summing total or average component scores. Such an approach assumes that the component tests have equal weighting, an assumption that needs to be verified if the resulting summary score is to reflect the relative importance of the component tests. To estimate the relative prominence of the LCA and LCG Listening and Reading tests, McDonald's Omega reliability was used to estimate, via confirmatory factor analysis (CFA) loadings, the relative weighting of the two component tests (see Hayes & Coutts, 2020). Table 4 below reports the results.

Table 4: CFA Standardised loadings of LCA and LCG tests

Test	Standardized loading
LCA_L	0.943
LCA_R	0.942
Test	Standardized loading
LCG_L	0.957
LCG_R	0.957

It can be seen that the two LCA and the two LCG tests have near equal loadings, indicating equal prominence. In such a case, summing up or averaging listening and reading in LCA and LCG to derive overall scores is justified. It is recommended that averaging be used to keep overall scores within the 100-point Global Scale. Using averages involves computing the mean of the component tests – two in the case of LCA Listening and Reading in Table 4 above. If the component tests do not have near equal weights, the mean would advantage the test/s with lower weight and disadvantage those with high weights, resulting in inaccuracies in the overall scores and leading to, in extreme cases, possible candidate appeals.

External Triangulation: Comparison with IELTS

In order to establish the extent to which results on the LCA and LCG related to those on another internationally used exam, around 500 candidates took one of the two LanguageCert tests (LCA or LCG) and the IELTS equivalent.

The statistical procedure appropriate in such situations is the multinomial test. This test estimates the equivalence of CEFR levels obtained by candidates in the LCA/LCG tests and the CEFR levels obtained on the IELTS test.

The Bayesian version of multinomial tests further estimates the range of variation in the two sets of rankings in the population, known as the 95% credible interval (CI). If and when the two sets of rankings fall within the CI, they are deemed to be equivalent. Understandably, the multinomial test is sample size sensitive. Given the relatively small sample size, six sets of model-based simulated data were generated for the LCA and LCG data, extending the initial total sample of 500 for both tests to a large sample of over 3,000. Bayesian multinomial test credible intervals (CI) were calculated for the comparative distributions to provide an indication of future distributions. Table 5 presents the results.

Table 5: LCA/LCG and IELTS comparative distributions

CEFR	LCA/LCG	IELTS	95% Credible Interval (CI)	
			Lower	Upper
A2	1%	1%	0%	1%
B1	8%	8%	7%	9%
B2	32%	32%	30%	33%
C1	30%	29%	28%	31%
C2	30%	30%	29%	32%

Both projected LCA/LCG as well as IELTS sample totals were within lower-upper CI ranges. To exemplify, it was projected that 30% of the LCA/LCG and 29% of the IELTS sample would obtain C1. The 95% credible interval for the percent obtaining C1 was projected to be between a lower bound of 28% and an upper bound of 31%, which was indeed the case. These are very compelling findings.

Global Scale / Raw Score Conversions

With the Global Scale in place, it is now possible to produce a range of indicators or metrics which relate directly to the GS. One of these, which will need to be test-specific, is the concept of the raw score conversion table. For a given test, the raw score conversion table maps the raw score to the Rasch-calibrated 100-point scale. Table 6 presents a sample of the LCA test mapped on to the Global Scale.

Table 6: LCA test mapped on to the Global Scale

Score	Listening	Reading	
30	100	95.81	
29	100	95.47	
28	98.38	92.16	
27	92.41	91.29	
26	89.71	90.7	C2
25	84.53	86.13	
24	81.24	82.47	
23	80.92	79.01	
22	80.4	77.28	
21	78.07	77.04	
20	75.53	76.75	
19	74.84	76.73	
18	73.57	76	
17	73.12	75.78	
16	72.54	74.32	C1
15	69.76	73.32	
14	68.36	72.96	
13	67.65	72.42	
12	65.03	71.88	
11	64.78	70.6	
10	62.29	69.68	
9	59.33	69.65	
8	57.48	64.14	
7	57.46	61.13	
6	56.33	60.08	B2
5	55.48	58.32	
4	51.36	55.46	
3	50.6	55.01	
2	46.54	52.44	B1
1	40.95	41.49	A2

While score correspondences between skills may be expected to be close, they will not necessarily be exactly the same. This may be seen in Table 6 above. At C2, both Listening and Reading have a lower bound cut score at the same point. At C1 and below, however, Reading is somewhat offset to Listening with slightly lower cut score points. However, while some variance between test forms is inevitable, it is important to reduce such variance to a minimum, and it is in this context that the methodology described here is important.

Global Scale Score Report

The Global Scale allows ease of interpretation for test users and a finely tuned results service across all language skills. As shown, performance can be separated in each skill and overall, so that a test taker is not only described as having 'C1 ability', for example, but a more precise level of detail is provided on test taker's performance. The Score Report shows an overall score, the overall CEFR level of attainment reached, and the score for each of the skills using both the Global scale and the CEFR level of attainment. Appendix 1 presents a sample of a certificate for LanguageCert Academic reporting Global Scale scores.

In Closing

This paper traces the development of the LanguageCert Global Scale from the original LID scale. The process began with the establishment of a set of Rasch-calibrated item locations for the LanguageCert Test of English (LTE) test items. The LID scale was then calibrated, and the precision and stability of the scale established on the basis of overall reliability and construct validity. The LID scale was found to be sufficiently robust and after calibration aligned well to the Global Scale with an appropriate mean and logit value. The Global Scale was then used to calibrate and map the LCA and LCG Listening and Reading tests. An alignment with IELTS-based CEFR levels of candidates in the LCA and LCG tests with CEFR levels specified within the Global Scale resulted in a remarkably close match.


The LanguageCert Global Scale may, it can be seen, be taken as appropriately established with a strong developmental background and rigorous validation procedures. External cross validation established via correspondences with IELTS underscores the robustness of the LanguageCert Global Scale, illustrating its clear links to the CEFR.

To conclude, the detail outlined about the development of the LanguageCert Global Scale illustrates how the Scale forms a solid foundation for all LanguageCert tests to expand into a language test pool to assess most language ability areas with good assessment quality and a stable standard mapped to the CEFR.

References

- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Milton Park, UK: 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.
- Handy, C. (1995), *The Sigmoid Curve in The empty raincoat: Making sense of the future*, Arrow Books, London, pp. 50–57.
- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *But.... Communication Methods and Measures*, 14(1), 1-24.
- Jones, C. (2023). An exercise in evolution: refocusing LanguageCert IESOL C1 to the academic context. In Falvey, P., & Coniam, D. (eds.) (2023). *Certifying quality in assessment and learning: Research and validation at LanguageCert*. Volume 2. London, UK: LanguageCert.


Appendix 1: Certificate Reporting LanguageCert Academic Global Scale Scores



LanguageCert Academic
(Listening, Reading, Writing, Speaking)

Test Report

Candidate Information


Last Name:	Candidate's Last Name	
First Name:	Candidate's First Name	
Date of Birth:	xx Month xxxx	
Candidate Number:	99800...	
UKVI Candidate URN:	PPC/...	
ID Type:		
ID Number:		Nationality:

Test Centre Information

Date of Test:	xx Month xxxx	Date Test Results Issued:	xx Month xxxx
Test Centre number:		Test Centre country:	
Mode of Delivery:			

Candidate Results (out of 100 on the LanguageCert Global Scale)

Listening		Writing	
Reading		Speaking	
Total Score			
CEFR Level			



Marios Molfetas
LanguageCert
Responsible Officer

THIS IS NOT A CERTIFICATE

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926.
LanguageCert reserves the right to amend the information given before issuing certificates to successful candidates.
info@languagecert.org

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926.

Copyright © 2023 LanguageCert

All rights reserved. No part of this publication may be reproduced or transmitted in any form and by any means (electronic, photocopying, recording or otherwise) except as permitted in writing by LanguageCert. Enquiries for permission to reproduce, transmit or use for any purpose this material should be directed to LanguageCert.

DISCLAIMER

This publication is designed to provide helpful information to the reader. Although care has been taken by LanguageCert in the preparation of this publication, no representation or warranty (express or implied) is given by LanguageCert with respect as to the completeness, accuracy, reliability, suitability or availability of the information contained within it and neither shall LanguageCert be responsible or liable for any loss or damage whatsoever (including but not limited to, special, indirect, consequential) arising or resulting from information, instructions or advice contained within this publication.



Language
Cert

languagecert.org