LANGUAGECERT®

From Displacement to Stability: Iterative Anchoring in Rasch Calibration of the LTE Adaptive Reading and Listening Tests

David Coniam
Tony Lee
Leda Lampropoulou

October 2025

From Displacement to Stability: Iterative Anchoring in Rasch Calibration of the LTE Adaptive Reading and Listening Tests

David Coniam, Tony Lee, and Leda Lampropoulou

Abstract

This paper reports an operational iterative anchoring procedure for equating items in the LANGUAGECERT Test of English adaptive Reading and Listening tests. Using expert-judged anchor values (Reading: 76; Listening: 36) across large item pools (Reading: 489; Listening: 442), we calibrated with Winsteps software and enforced a displacement rule: anchors with displacement greater than 0.5 logits were released and the scale re-estimated. Both skills stabilised after four iterations; 27 of 76 Reading anchors (35.5%) and 10 of 36 Listening anchors (27.8%) remained, with very satisfactory Rasch fit and reliability statistics confirming robustness. Anchor attrition reflects quality control rather than instability: only the most stable items are retained to preserve the measurement frame. This process highlights the practical value of iterative anchoring in large-scale adaptive testing programmes.

Keywords: Rasch measurement, displacement, reading and listening tests, adaptive test

Suggested citation

Coniam, D., Lee, T., & Lampropoulou, L. (2025). *From Displacement to Stability: Iterative Anchoring in Rasch Calibration of the LTE Adaptive Reading and Listening Tests*. LANGUAGECERT.

Rasch Common Item Anchoring

Equating two or more test forms to an existing measurement scale requires linkages between tests. In Rasch measurement, such linkages may be achieved via either common persons or common items, although the modus operandi is the latter, namely via anchor items (Aryadoust et al., 2020). Calibration involves estimating the locations of a set of items and a sample of persons on the probabilistic Rasch scale (Linacre, 2024), yielding a *frame of reference* (FOR) specific to the items, persons, and data used (Humphry & Andrich, 2008). Rasch (1977) referred to this property within Rasch measurement as *specific objectivity*, meaning that calibrated values of items are contextualised and objective within a particular FOR, but equating across tests requires stable anchors to preserve that objectivity.

The shifting of items in a test during the process of equating refers to the test as a whole (see Lee et al., 2022). If there is a common measurement scale involved, this serves as an underlying measurement scale not unlike the metre. The common measurement scale should not then be taken as an actual scale consisting of items. The tests mapping onto a common measurement scale have a common reference metric but need to remain distinct (Goodman, 1990).

In the above context, irrespective of how a calibration may be manipulated, the FOR – the ordering of items and persons resulting from the calibration – needs to be preserved (see e.g., Coniam et al., 2022). In the current paper, anchoring via common items is described, with the anchoring process undergoing as many iterations as necessary to resolve unacceptable displacements (see Linacre, 2024: 144). The displacement statistic indicates how much an item's difficulty estimates shifts when it is treated as an anchor compared to when it is freely estimated (Stahl and Muckle, 2007).

Displacement is important since it reports the degree of match between the calibration of anchor items and original item locations, and can reveal misfit between the anchor items and the current sample. The set of anchor items selected for the test linking / equating after each iteration are therefore chosen such that the overall calibration results are not interfered with.

A small displacement typically suggests that the anchor item is functioning consistently, supporting its use for establishing a stable scale.

In contrast, a high displacement value – conventionally over half a logit (Wright & Douglas, 1976) – indicates that an item's difficulty estimate may not be stable across samples or conditions. It may indicate issues such as possible misalignment between an item's difficulty and the ability of the sample.

Since anchor items with high displacement values may unduly disturb the original calibration, they should then be unanchored, and a further calibration linking / equating conducted. Linacre (2024) states:

> *items with displacements of more than 0.5 logits, that are also bigger than the item S.E.s, are candidates for recalibration.*

The recalibration process may involve a number of iterations until all displacement values are, ideally, near zero. A further issue relates to how many anchor items are needed in the calibration - recalibration process. Linacre (2024) suggests that:

> *The percent of anchor items is less important than the number of anchor items. We need enough anchor items to be statistically certain that a large incorrect value of one anchor item will not distort the equating.*

While the concept of displacement is not new, there has been minimal discussion in the literature of how to operationalise it in practice. To achieve a steady state whereby no items have displacement values greater than the 0.5 logit maximum recommended may well require a number of iterations. The process is referred to in this paper as *iterative anchoring*.

In light of the agendas laid out above, the current study contributes by documenting the process of iterative anchoring in a large-scale operational setting. Specifically, we report on the LANGUAGECERT Test of English (LTE) adaptive Reading and Listening tests, administered in 2023–2024. Each test involved over 400 items and a comparatively large set of expert-judged anchors (76 in Reading, 36 in Listening). In essence, the paper illustrates how iterative anchoring ensures robust calibration in large adaptive language assessments, and how apparent anchor loss is an expected outcome of a deliberate stress-testing process.

LANGUAGECERT Test of English (LTE) Adaptive Test

An analysis of two tests – a Reading test and a Listening test – is described below. These were drawn from the LANGUAGECERT Test of English (LTE) adaptive test administered in 2023-2024.

The LTE adaptive test is a level-agnostic assessment of listening and reading which reports test-taker results from CEFR levels A1 to C2. The LTE draws from a series of item banks, each comprising approximately 1,000 items that include a range of listening and reading items and testlets of between two and five items.

The Listening test component comprises four sections targeting different reading sub-skills. These include understanding detailed information, following the sequential aspects of an exchange, comprehending longer spoken texts and, at the higher CEFR levels, appreciating speaker intention, inference, and summarising.

The Reading Test component comprises six sections, each focusing on different reading sub-skills. These include reading and understanding short notices, vocabulary use in context, lexico-grammatical awareness, and comprehension of longer texts which tap into different reading sub-skills depending on the level of the test taker (from understanding factual information at lower levels to recognising writer intention at higher levels.)

The LANGUAGECERT adaptive test algorithm functions such that all test takers are presented with 58 items – normally 28 listening and 30 reading items. The first item presented is at approximately B1 level. Test takers subsequently move dynamically between item types depending upon performance and predicted ability.

After having completed the 58 items, a test taker's level is determined at a specific point on the 100-point LANGUAGECERT Global Scale and its mapped CEFR level from A1 to C2. The data for this study is drawn from a dataset of more than one hundred thousand test takers who completed the LTE adaptive test in 2023-2024, both during pre-testing and live testing modes. It should be noted in passing that the LTE and its item bank have been calibrated jointly such that test takers receive the same grade whether they take the paper-based or the computer-adaptive version of the test.

Statistical Analysis and Baseline Descriptives

In the analysis reported below, calibrations were conducted via the software Winsteps (Linacre, 2024). The data matrix was calibrated with logit values rescaled to the LANGUAGECERT Item Difficulty (LID) scale mid-point of 100 and a spacing factor of 20 (see Pike et al., 2024). Table 1 provides detail on the number of items and anchors in the two tests – Reading and Listening – analysed in the current study.

Table 1: *Items and anchors in the LTE Reading and Listening tests*

Skill	Items	Anchors
Reading	489	76
Listening	442	36

The Reading test contained slightly more items and anchor items than the Listening test. Both tests, however, included large enough sets of anchor items to allow for their removal if displacement figures so indicated, while maintaining sufficient items for calibration purposes. For ease of exposition, most of the following discussion focuses on the Reading test, although the performance of the Listening test is also outlined. Before proceeding to an analysis of the iteration process, baseline descriptive statistics for the Reading test are provided in Table 2.

Table 2: *Baseline Descriptive Statistics*

	S.E.	Infit mean squares	Outfit mean squares
N	489	489	489
Mean	2.04	0.99	0.99
Std. Deviation	1.21	0.08	0.15
Maximum	11.8	1.33	2.05
99th percentile	6.29	1.2	1.45
75th percentile	2.24	1.04	1.07
50th percentile	1.85	0.98	0.98
25th percentile	1.35	0.94	0.91
Minimum	0.54	0.75	0.47

Infit and outfit mean square values show that very few items fell outside the 0.5–1.5 acceptable range (Lunz and Stahl, 1990). More significantly, at the 99th percentile,

standard errors were only just over half a logit (6.29 LID scale points), indicating that the data used in the calibration can be considered statistically robust.

Iterative Anchoring in Practice

The general procedure for iterative anchoring is as follows. A 'free' analysis – that is, one in which anchor values are not used – is first performed to verify that the new data are correct and contain no item miskeys, data-entry errors, miscoding, etc. This initial calibration is referred to as 'Iteration 0'.

Subsequently, a series of item-anchored analyses is conducted in which item displacements indicate the extent to which anchor values have 'drifted'. If any anchor item has a displacement greater than 0.5 logits, the item is unanchored and its value allowed to float freely along with all other items in the dataset (see Linacre, 2024).

The first analysis using anchor item values is termed 'Iteration 1'. Iterations continue until none of the anchor items display displacements greater than 0.5. Once this 'steady state' has been reached, the final item-anchored analysis is used for reporting reliability indices, fit statistics and other calibration metrics.

Research Questions

Two research questions are being pursued in this paper.

1. How many iterations are required for the set of anchor items to reach a 'steady state', defined as no anchor item having a displacement greater than 0.5 logits?
2. How many, or what percentage of, anchor items remain at the final iteration once this steady state has been achieved?

Iterations with the LTE Reading Test

In the discussion below, different facets of the iterative anchoring procedure are elaborated upon. The Reading test is the major focus in the analysis and discussion below since it contained slightly more items (489) than the Listening test as well as more anchor items to which expert-defined values had been assigned (76 against 36).

To arrive at a position whereby none of the 76 anchor value items had displacement values greater than 10 LID scale points, or half a logit, a number of iterations were required, as will be outlined.

For ease of interpretation, the tables below present the results for a limited set of 12 items (the first 12 items in the Reading test, as it happens) initially defined as anchor items. For these 12 items, the respective displacement and calibrated LID scale values are provided after each iteration. Table 3 presents the results of the initial iteration (Iteration 0). Here, it will be recalled, no anchor values were used; rather, items were allowed to float freely.

Table 3: *Iterative Anchoring – Iteration 0*

Item	Displacement 0	LID 0
R001	-0.1	120
R002	-0.1	123
R003	-0.1	125
R004	-0.1	117
R005	-0.1	120
R006	0	148
R007	0	170
R008	0	160
R009	0	160
R010	0	170
R011	-0.15	119
R012	-0.17	119

Displacement values can be seen in column 2 and LID values in column 3. Because the analysis is permitted to run as many as times as is necessary to converge, displacement values at this initial stage were small. No displacement exceeded five LID scale points, i.e., a quarter of a logit.

Next, the first actual anchoring iteration (Iteration 1) used the anchor values supplied for the 76 Reading test anchor items.

Table 4 below provides detail on the same 12 Reading test anchor items shown in Table 3 above. The difference this time is that anchor values were used, as indicated in Column 2, by the letter 'A' standing for 'anchor'. Column 2 states which items were set as anchor items (all 12, as can be seen). Column 3 lists displacement values, with those above half a logit (10 LID scale points) highlighted in red. Column 4 reports LID scale values incorporating displacement values.

Table 4: *Anchoring Iteration 1*

Item	Anchor 1	Displacement values 1	LID scale values 1
R001	A	-5.44	120
R002	A	6.7	123
R003	A	-26.63	93.28
R004	A	-11.8	99.91
R005	A	8.69	120
R006	A	20.71	170.67
R007	A	-7.9	170
R008	A	19.41	181.54
R009	A	4.27	160
R010	A	-1.01	170
R011	A	22.35	136.04
R012	A	0.04	119

As can be seen, five of the 12 items (i.e., those in red) had displacement values above half a logit. These items are subsequently unanchored and allowed to float freely. Table

5 below presents an update to Table 4. The final column in Table 5 ("→ Anchor 2") shows which items were retained as anchors for the next iteration.

Table 5: *Anchoring Iteration 1 with Anchor 2s defined*

Item	Anchor 1	Displacement values 1	LID scale values 1	→ Anchor 2
R001	A	-5.44	120	A
R002	A	6.7	123	A
R003	A	-26.63	93.28	
R004	A	-11.8	99.91	
R005	A	8.69	120	A
R006	A	20.71	170.67	
R007	A	-7.9	170	A
R008	A	19.41	181.54	
R009	A	4.27	160	A
R010	A	-1.01	170	A
R011	A	22.35	136.04	
R012	A	0.04	119	A

Examination of the whole 489-item dataset revealed that 46 anchor items had unacceptably high displacement values. Consequently, the number of anchor items was reduced from 76 to 30. While unachoring over half of the original set of anchors may appear drastic, Rasch analyses often converge differently with varying items, and different items may subsequently present as unacceptable. Table 6 shows how anchoring Iteration 1 extended into Iteration 2 using the same 12 items.

Table 6: *Anchoring in Iteration 2*

	Iteration 1				Iteration 2			
Item	Displace 1	LID 1	→ Anchor 2		Displace 2	LID 2	Anchor 2	→ Anchor 3
R001	-5.44	120	A		-10.86	107.89	A	
R002	6.7	123	A		1.44	123	A	A
R003	-26.63	93.28			-0.01	91.93		
R004	-11.8	99.91			-0.01	98.57		
R005	8.69	120	A		3.42	120	A	A
R006	20.71	170.67			0	170.62		
R007	-7.9	170	A		-6.04	170	A	A
R008	19.41	181.54			0	181.49		
R009	4.27	160	A		6.16	160	A	A
R010	-1.01	170	A		0.96	170	A	A
R011	22.35	136.04			-0.03	134.63		
R012	0.04	119	A		-5.73	119	A	A

As can be seen, the five items (in red) from Iteration 1 which were allowed to float had, by Iteration 2, settled into a steady state of acceptability. In Iteration 2, however, item

R001 reported a displacement value above half a logit, meaning that it needed to be unanchored for the subsequent Iteration 3. The final column ("→ Anchor 3") shows which items were retained as anchors for the following iteration, Iteration 3.

Across the full dataset, Iteration 2 showed that of the remaining 30 anchor items, only two showed unacceptable displacement values, resulting in the number of anchor items being reduced to 28.

In Iteration 3, one anchor item again showed a displacement above 0.5 logits and was unanchored, reducing the set to 27.

It Iteration 4, no anchor items reported displacements above the 0.5-logit threshold. At this point, the iteration process was therefore terminated.

To illustrate the iteration process visually, a schematic summary is provided in the Appendix. The figure shows how, after the first major iteration, a substantial adjustment – or "shakeout" - occurs, while subsequent iterations show only minimal further change. This pattern indicates that the anchoring process stabilised quickly after the initial iteration, with later iterations confirming the attainment of a steady state.

The same iteration process was repeated with the LTE Listening test items. The results are summarised below in Table 7, which provides a comparative picture of the calibration for both the Reading and Listening items.

Table 7: *Calibration of the Reading and Listening items*

Skill	Items	Anchors	Anchors removed after Iter^{n} 1	Anchors removed after Iter^{n} 2	Anchors removed after Iter^{n} 3	Anchors removed after Iter^{n} 4	Iterations Needed to full calibration	Good anchors
Reading	489	76	46	2	1	0	4	27/76
Listening	442	36	20	4	2	0	4	10/36

As shown, both Reading and Listening tests required four iterations for the items to be acceptably calibrated; by Iteration 4 no item displayed a displacement value greater than half a logit.

Discussion

This study has reported on the process of item calibration using anchor items within the LTE adaptive Reading and Listening tests administered in 2023-2024. The investigation focused on determining the number of iterations required for the set of anchor items to reach a 'steady state' – defined as no anchor item showing a displacement value greater than 0.5 logits. A secondary aim was to establish the number or the percentage of anchor items remained once this steady state had been achieved.

The LTE adaptive Reading and Listening tests were selected for analysis because both featured large item pools – each containing over 400 items – and relatively substantial sets of anchor items (76 for the Reading and 36 for the Listening test). These large

numbers provided a suitable basis for exploring how successive iterations function in practice.

For the Reading test, four iterations were required, at the end of which 27 out of the 76(35.5%) of anchor items remained. A similar figure was recorded with Listening where 10 of the 36 (27.8%) anchor items remained after four iterations. Whether such percentages are typical or low will require future research. Nevertheless, the close alignment between the two skills, together with previous evidence that LANGUAGECERT's expert-defined anchor values are reliable (Coniam et al., 2024), supports the robustness of the present calibration process.

In addressing the question of the number of anchor items necessary for effective equating, the findings indicate that quality outweighs quantity. In other words, it is not simply the case that more anchor items produce better equating; rather the acceptability of the anchors depend on smaller displacement statistics. Pibal and Cesnik (2019: 3) state:

> *… while a higher number of anchors might help obtain a better estimate of the common scale, it is not always the quantity but rather the quality of the anchors that is decisive in tying different scales together.*

The current analysis reinforces this view. Sufficient anchor items to be provided ab initio to enable several iterations, but only those demonstrating stability should remain. As long as a sufficient number of acceptable anchor items remain at the end of the series of iterations across the range of item difficulties, the resulting calibration can be regarded as robust and defensible.

Conclusion

The study demonstrates that iterative anchoring offers an effective, evidence-based approach to test equating within adaptive language assessments. By progressively identifying and removing unstable anchors, the process safeguards the integrity of the Rasch frame of reference and ensures that only stable items contribute to the common measurement scale.

The results underline two key principles. First, the iterative process is not a sign of instability but a form of systematic quality control. Second, the goal of anchoring is to align the *test as a whole* rather than to preserve specific item parameters. The final anchor set should span the full range of item difficulties while maintaining the overall person–item distribution pattern.

In sum, iterative anchoring enables LANGUAGECERT to maintain a fair, transparent, and statistically sound calibration framework, ensuring that test scores remain directly comparable across administrations and delivery modes.

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Coniam, D., Lee, T., & Lampropoulou, L. (2024) The use of expert judgement and pairwise comparison in the establishment of an assessment scale. *Journal of Applied Linguistics, 37*. https://doi.org/10.26262/jal.v0i37.10392

Coniam, D., Lee, T., Milanovic, M., Pike, N., & Zhao, W. (2022). Role of expert judgement in language test validation. *Language Education & Assessment, 5*(1), 18-33. https://doi.org/10.29140/lea.v5n1.769

Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. *Sociological Methodology, 20*, 249-294. http://doi.org/10.2307/271088

Humphry, S. M., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement, 9*(3), 249-264.

Lee, T., Milanovic, M., & Pike, N. (2022). Equating Rasch values and expert judgement through externally-referenced anchoring. *International Journal of TESOL Studies, 4*(1), 187-202. http://doi.org/10.46451/ijts.2022.01.12

Linacre, J. M. (2024). *Winsteps® Rasch measurement computer program User's Guide.* Version 5.8.3. Portland, Oregon: Winsteps.com.

Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*(4) 425-444. https://doi.org/10.1177/0163278790013004

Pibal, F., & Cesnik, H. S. (2019). Evaluating the quantity-quality trade-off in the selection of anchor items: A vertical scaling approach. *Practical Assessment, Research, and Evaluation, 16*(1), 6. doi: https://doi.org/10.7275/nncy-ew26

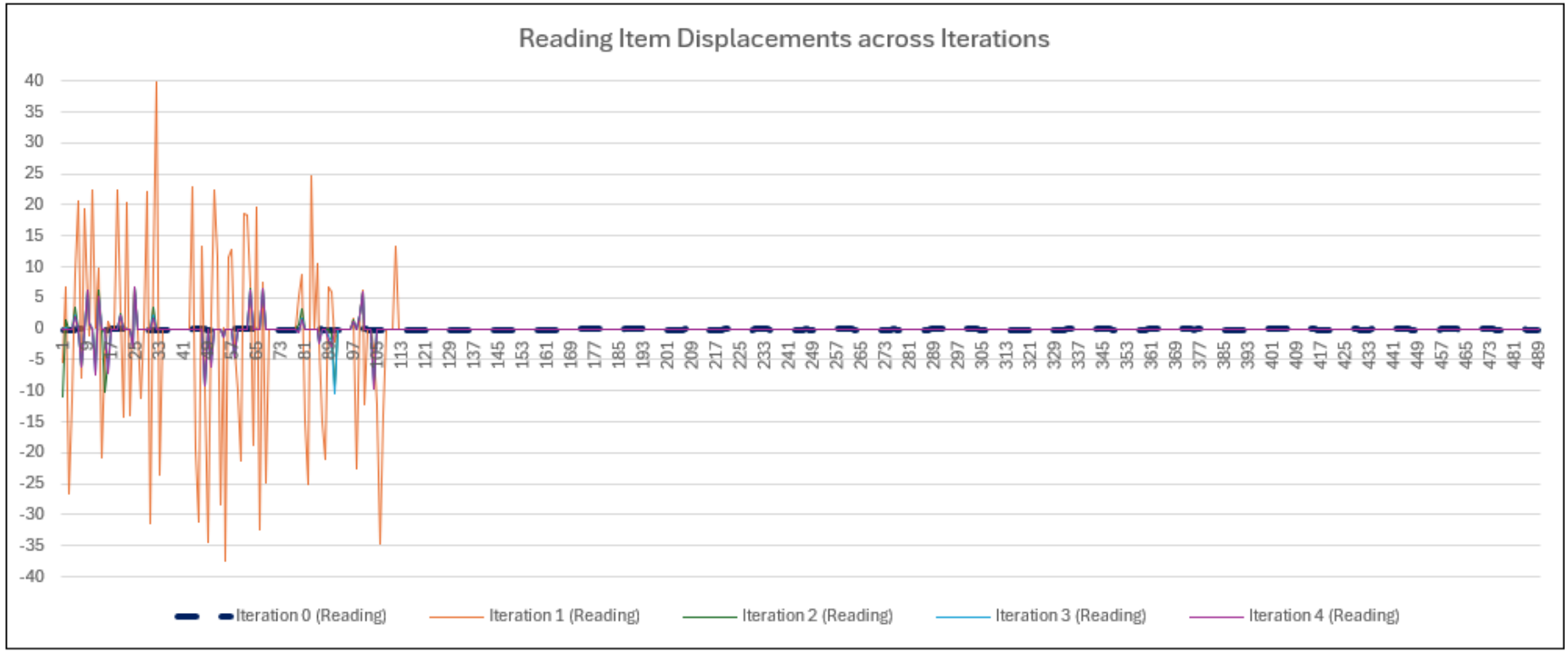
Pike, N., Papargyris, Y., Dourda, C., Lee, T., & Coniam, D. (2024). *Recalibrating and extending the analysis of the LANGUAGECERT Test of English*. London, UK: LANGUAGECERT.

Rasch, G. (1977). On Specific Objectivity an Attempt at Formalizing the Request for Generality and validity of Scientific Statements. *Danish Yearbook of Philosophy, 14*(1), 58-94. https://doi.org/10.1163/24689300-01401006

Stahl, J., & Muckle, T. (2007). Investigating drift displacement in Rasch item calibrations. *Rasch Measurement Transactions, 21*(3), 1126-1127.

Wright, B. D., & Douglas, G. A. (1976). Rasch item analysis by hand. *Research Memorandum 21*. Chicago, IL: University of Chicago.

Appendix: Schematic representation of the four iterations



The figure provides a schematic representation of the four iterations for the LTE Reading Test. The 76 anchor items were positioned among the first 113 items.

Iteration 0 is represented by dark blue markers along the horizontal zero line across all 489 items. **Iteration 1** (orange lines) shows variations extending up to +40 (two logits) and –30 (1.5 logits) LID scale points. **Iteration 2** (green lines) shows much smaller variations, remaining within ±10 LID scale points. **Iteration 3** (blue lines) also remains within ±10 LID scale points. **Iteration 4** (purple lines) similarly remains within ±10 LID scale points. After the first major iteration, a noticeable adjustment occurs, while subsequent iterations show minimal movement. This demonstrates that the process quickly converged to a stable configuration.

LANGUAGECERT®

by PeopleCert

MILLIONS OF EXAMS DELIVERED WORLDWIDE

LANGUAGECERT is an Awarding Organisation recognised by Ofqual. It spearheads innovations in language assessment and certfication, providing high-quality services to the global learners' community. It is a UK-based member of PeopleCert Group, a global leader in the certification industry, that delivers millions of exams in over 200 countries.

Learn more about LANGUAGECERT exams at:
www.languagecert.org

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

info@languagecert.org

/LanguageCert.org