Australian Government
Department of Foreign Affairs and Trade

ACER

WG/GAML/12

Pairwise Comparison Method

Concept Note
November 2022

The Global Education Monitoring (GEM) Centre drives improvements in learning by supporting the monitoring of educational outcomes worldwide. The GEM Centre is a long-term partnership between the Australian Council for Educational Research (ACER) and the Australian Government's Department of Foreign Affairs and Trade (DFAT).

Global Alliance to Monitor Learning meeting
23 November 2022

unesco
Institute for Statistics

GLOBAL ALLIANCE TO MONITOR LEARNING

Contents

Introduction

Sustainable Development Goal (SDG) 4 aims to ensure that, by 2030, "all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes."

Indicator 4.1.1 refers to the proficiency indicator referring to three levels of schooling: end of lower primary, end of primary, and end of lower secondary and two subjects (reading and mathematics). The indicator reads as follows:

"4.1.1 Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level [MPL] in (i) reading and (ii) mathematics, by sex."

Montoya (2022) sets out that the reporting format of the indicator aims to communicate two pieces of information:

1. the percentage of students meeting at least minimum proficiency standards for the relevant domains (mathematics and reading) for each point of measurement (grades 2/3; end of primary and end of lower secondary) and

2. whether a program can be considered comparable, and the conditions under which the percentage of children at or above MPL can be considered comparable to the percentage reported from another country

She also states that the indicator needs the following inputs:

- **Domain**: reading and mathematics. Reading and mathematics are measured at the national level in numerous way
- **Minimum proficiency level (MPL)**: is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) at a given age/grade
- **Linking to the MPL**: methodologies to harmonize various data sources to a common definition of the MP
- **Sample**: the sample needs to be representative of the relevant population

The paper sets out several approaches, both statistical and non-statistical, by which jurisdictions may choose to link their assessments to the global standards for reporting. One of these, the Pairwise Comparison Method (PCM)[1], was first set out in Lazendic (2019). This concept note provides further details on the PCM and the work that has taken place since 2019 to enable the approach to be operationalised.

[1] Also known as the Comparative Judgement method.

Learning Progression Scales

The PCM relies on the Learning Progression Scales (LPSs) for reading and mathematics, developed by the Australian Council for Educational Research (ACER). A full description of the development of the scales is provided in the section Technical Note on the construction of the Learning Progression Scales, but in essence they are robust, statistical ordering of items drawn from a range of different assessments that was developed using a pairwise comparison approach. The International Standard Setting Exercise (ISSE) was carried out by ACER in 2022 to establish the MPL threshold on the LPSs for reading and mathematics (ACER, 2022). The ISSE used the Bookmark method to set the threshold, involving participants with a diverse range of experience, background and skill, including sufficient geographical representation.

Overview of the Pairwise Comparison Method

As with all approaches to linking to global standards, the first stage of the PCM is to determine whether the assessment instrument is of sufficient validity to be suitable for SDG 4.1.1 reporting. For the PCM, this will be achieved using the same self-assessment exercise that is currently being updated for the policy linking toolkit. This will require jurisdictions to consider their assessment against a set of criteria:

- **MPL alignment** – does the content of the assessment align with the MPL, determined by comparison with the domains, constructs and subconstructs of the Global Proficiency Framework (GPF)?
- **Item review** – were the items in the assessment been reviewed quantitatively and qualitatively to determine their appropriateness for inclusion in the assessment?
- **Sample** – was the assessment administered to a cohort that is representative of the population against which the results would be reported?
- **Administration** – was the assessment administered in a standardised way?
- **Reliability** – were the outcomes of the assessment sufficiently reliable?

If the assessment is determined to be sufficiently valid, the jurisdiction can set up a panel to undertake a PCM exercise with items from their assessment and a selection of items from the relevant LPS.

Pairwise comparison methods exploit the finding that people are better at comparing two objects or examples of student work against each other, than at evaluating one object or piece of student work against criteria (Thurstone, 1927). Based on multiple comparative judgements, a rank order of tasks or examples of student work is generated. This rank order is based on all decisions made across judges, and results in reliable scales.

The advantage of pairwise comparisons is that there is no limit regarding the type and form of the assessment tasks. Furthermore, a large number of items can be included in the pairwise comparison as this judgement process is efficient. Thus, this method can provide robust and reliable empirical linking with MPLs. The robustness of pairwise scaling and statistical linking can be analysed and evaluated using standard IRT fit and reliability statistics.

The judgement panel can easily be run remotely, enabling a greater diversity of panellists to be involved, and ACER has developed its award-winning software platform, Signum[2], to facilitate the exercise. Panellists are shown pairs of items from the pool containing items drawn from the LPS and those from the assessment being linked and are asked to determine which is more difficult. At the end of the process, the items from the assessment under judgement are embedded within the LPS. This allows the MPL thresholds determined within the ISSE to be applied to the assessment and enables reporting against SDG 4.1.1.

Advantages of the Pairwise Comparison Method

The PCM has a number of advantages over other methods that can be used to report against SDG 4.1.1:

- It is cheaper to run than statistical linking methods, providing a cost-effective way for jurisdictions and development partners to quickly align an assessment to global standards
- The task for panellists is relatively simple (which item is more difficult) and does not require them to internalise the standards in the MPLs/GPF or determine whether pupils at different standards would have a two-thirds chance of answering the item correctly, which can be difficult for teachers
- The exercise can be run entirely remotely using a specifically designed platform to maintain consistency
- As new assessments are subject to the PCM, they can be added to the LPS (if security requirements allow) to build an invaluable resource to support capacity development

Next steps

ACER is developing a toolkit to enable the consistent implementation of the PCM and is looking for jurisdictions in which the approach can be implemented.

At present, the LPS is only available for use with assessment items in English; however, ACER would be interested to implement the approach with a bilingual panel to determine

[2] https://www.acer.org/gb/discover/article/acer-at-the-eassessment-awards-2022

if it is possible to link items in another language to the same LPS or whether separate LPSs are required for each language.

Technical note on construction of the Learning Progression Scale

The purpose of the technical note is to provide the overview of the Learning Progression scale devolvement and a set of research studies conducted to collect empirical evidence for the Pairwise Comparisons Method feasibility including robustness of the prosed statistical linking approach.

The Learning Progression Scales have their origin in work done in collaboration with UNESCO Institute for Statistics (UIS) to develop empirically based, described reporting scales for reading and mathematics learning areas. The ISSE is a step towards the long-term goal of locating Proficiency Levels on these scales associated with the end of lower primary, end of primary and end of secondary schooling, and demonstrating empirical and qualitative growth between these levels.

The earliest versions of these scales were developed to capture the range of reading and mathematics skills which students across the world in primary and secondary schools have demonstrated in international, regional and national assessment programs. To construct a single scale for each learning area (reading and mathematics), hundreds of test items from a wide range of assessment programs were ordered by difficulty using statistical methods and expert judgments. These ordered test items were analysed to identify the kinds of skills required to answer each of the items correctly. These skill descriptions form the basis of the Learning Progression Scales.

ACER developed qualitative descriptions to articulate the order of skill development, initially in reading and mathematics, and to exemplify the range of skills that can be demonstrated at a particular point on associated empirical scales. To integrate items from different assessments onto a single scale ACER used a pairwise comparison technique; this places items used in the development of the Learning Progression Scale on the same item difficulty scale. In the pairwise comparison exercise, pairs of items are shown to subject matter experts, who then decide, for each pair, which item is more difficult. To analyse these judgments and produce an item difficulty scale for items included in the exercise, the Bradley-Terry-Luce (BTL) model is used.

BTL model is functionally equivalent to one-parameter IRT model and thus BTL item difficulty scales have the same properties as those developed in assessment programs from which the items have been sourced. Consequently, the BTL item locations can be interpreted in the same way and used to develop a set of described proficiency levels using the same approach as used in large-scale assessments. ACER used these initial learning progression scales to extend and refine MPL descriptors and to develop 600 new items for inclusion in UIS' Global Item Bank, many of which target the MPLs for reading and mathematics from lower primary to end of lower secondary.

ACER developed a program of research to:

- demonstrate the feasibility off ability of statistical linking between LP scale and the existing assessment program
- evaluate the validity of pairwise comparison method to establish the robust measurement scale underpinning the initial learning progressions and support further MPL descriptor elaboration including targeted item development.
- assess robustness of the pairwise comparison method's operational deployment requirements and solution

Study 1: Evaluating the core premise of PCM statistical linking approach

The main purpose of the first study was to establish the extent to which the pairwise comparison method (PCM) can recover the item location estimated using student responses in an existing, well-established assessment program. This is the core premise of the statistical linking approach proposed in PCM. The goal was to show that the item difficulty parameters obtained in PCM meet the reliability criteria a standard IRT based assessment linking. The further purpose was to estimate the reliability of the BTL item difficulty locations across different Comparative Judgement (CJ) exercises and thus provide evidence for the overall PCM robustness.

Materials

A sample of items from ACER's Progressive Achievement Tests (PAT) representative of the learning progression scale underpinning it were included in Study 1. A total of 60 items from PAT reading and 61 of items from PAT mathematics tests were used. In addition, 50 reading and 60 mathematics items from the pool of items used to establish the initial learning progressions.

Design

A total of 2994 and 2780 pairs were constructed, respectively for mathematics and reading CJ. On average each item was compared to any other items 42 times with a maximum of 50 and minimum of 30 comparisons per item.

The distance of items in the pair was restricted to be no wider than four levels on the initial learning progressions or PAT scale level. While the levels of the two scales were not equivalent, they provided sufficient information to avoid most of the trivial comparison where a very easy item is paired with a very hard item.

The position of an item in a pair was randomised across allocated pairs. The pairs were randomly ordered with a condition that a single item could appear in up to three pairs sequence. This is done to reduce CJ task cognitive load for the judges. Each judge was then randomly assigned a set of pairs from the overall pair sequence.

Such pairs construction, sequencing and allocation is a standard approach in the CJ technique and is used in all studies described in this note.

Participants

Judges were recruited from the pool of ACER item writers and subject matter experts in mathematics and reading. They all had a good understanding of PAT assessments and the concept of learning progressions and their role in item construction and scale descriptor development. A total of 21 judges in reading and 20 judges received a full day training on the initial learning progressions including some practical activities as a part of another research study conducted immediately before Study 1.

Procedures

Judges received training on the PCM concept and procedure including the online system used to administer the pairs and collect student responses during a two-hour session. The judges were then provided access to the online platform and were provided an hour to engage in the PCM exercise at their own pace. Judges then completed the exercise on their own during the next 48 hours. The average number of judgments across judges was 132 for reading and 157 for mathematics

Results and discussion

The judges' responses were analysed using the BTL model and all items were scaled to respective PCM scales with a mean of zero and a standard deviation of one logit.

A strong correlation between the original PAT scale location and the PCM scale location was observed for PAT items in reading and mathematics domains; r=0.96 and r=0.93.

These results provide strong evidence that PCM can reliably recover the original PAT item locations based on the students' responses IRT analyses. Such an outcome confirms that PCM can produce the scale parameters of the same quality and reliability expected in the standard IRT linking.

Furthermore, for items used in the initial learning progression development study, similar high correlation was observed between item BTL model locations observed during the earliest comparative judgement (CJ) exercise and that observed in Study 1; r=0.97 for reading and r=0.95 for mathematics. The results show that the CJ technique underpinning the PCM yields highly reliable item difficulty estimates across different judges, item pools and time.

Study 2: PCM replication and robustness

The purpose of the second study was to replicate the recovery of PAT items' location observed in Study 1 using different pool of judges and different approach to judges training. A new set of judges were selected from the pool of markers without operational experience with PAT item development and marking. This situation corresponds to a PCM deployment situation where participants are likely to have minimal or no experience

with LPS items that will be used in PCM linking exercises. The amount of training was also reduced compared to that of Study 1 to further match the conditions of the PCM deployment.

Materials

The same sets of PAT items used in Study 1 were included in Study 2. A sample of items from Pacific Islands Literacy and Numeracy Assessment (PILNA) was also included in the study. PILNA item sample included 40 reading items from literacy tests and 62 items from mathematics test. Finally, a selection of items ACER constructed to target and exemplify the MPLs and develop MPLs descriptors (hereafter- MPL items) were also included in the study. Total of 143 and 147 MPL items for reading and mathematics respectively were used.

Design

A total of 5250 and 4334 pairs were constructed, respectively for mathematics and reading PCM. On average each item was compared to any other items 36 times with a maximum of 48 and minimum of 25 comparisons per item for mathematics and 35 times with a maximum of 48 and minimum of 20 comparisons per item for reading.

The same pair construction and allocation approach described in Study 1 was used with PILNA level used for PILNA item and targeted LPS level used for MPL items.

Participants

Judges were recruited from the pool of ACER item markers. The participants had no recent marking nor assessment construction experience with PAT assessments. A total of 16 judges were allocated 270 comparisons in reading domain and 15 judges were allocated 350 comparisons in mathematics.

Procedures

Judges received introduction to LPS training remotely by subject matter experts. Judges were gathered in a marking centre for this training. The training session was approximately three hours long, with a plenary session on the concepts behind the LPS and separate unpacking of the LPS by domain. Training on the PCM concept and procedure including the online system used to administer the pairs and collect judges responses was delivered in person in a one-hour session. The judges were provided access to the online platform and were provided with the rest of the afternoon to start the pairwise comparison exercise. Judges completed the exercise on their own during the next 48 hours.

Results and discussion

The judges' responses were analysed and scaled following the procedure described in Study 1. A very high correlation between the original PAT scale location and the PCM scale location was again observed in Study 2; r=0.93 and r=0.91 for reading and mathematics respectively. Furthermore, even higher corelation was observed between

PCM locations observed in Study 1 and Study 2; r=0.97 for reading and r=0.94 mathematics.

The correlation between PCM scale location and the original PILNA scale location was also high r=0.83 and r=0.76 for mathematics and reading respectively. PILNA literacy scales reading and writing items together and thus the somewhat lower correlation in reading might be due to this construct misalignment between PILNA and PCM scales. In addition, the PILNA items cover grade 4 and 6 and thus have reduced spread against the LPS compared to that of PAT items. These caveats aside the results show that high level of corelation can be obtained from assessments expected to be encountered in the PCM operational deployment.

Taken together with the results of Study 1, the results of Study 2 provide strong empirical evidence for reliability and robustness of the PCM scaling and approach.

Study 3: Operational deployment feasibility

The third study was conducted in collaboration with the Educational Quality and Assessment Programme (EQAP), which is the organisation responsible for PILNA. EQAP contributed subject matter experts for the CJ exercise and supported the study implementation in Fiji. Training for and monitoring of the PCM exercise was provided by ACER remotely from Australia. The purpose of the third study was twofold: to replicate PCM scale ordering of PILNA items using experts familiar with PILNA; and to place the remining MPL items on the PCM LPS scale.

Materials

The study included PILNA items used in Study 2 and also included several items from the COVID-19: Monitoring the Impacts on Learning Outcomes (MILO) study. MILO was conducted by ACER in collaboration with the UIS and the GEM centre and was funded by the Global Partnership for Education (GPE). The MILO study measured learning outcomes in six African countries, with the aim of identifying the impact of COVID-19 on learning and reporting on the level of attainment for the SDG Indicator 4.1.1b. Total of 170 reading and 150 mathematics MPL items was used in the study. The number of PAT items was reduced to keep the number of comparisons manageable.

Design

The design was the same as that used in Study 2 with a total of 3877 and 2400 pairs constructed, respectively for mathematics and reading. On average each item was compared to any other items 28 times with a maximum of 44 and minimum of 14 comparisons per item for mathematics and average of 19 times with a maximum of 30 and minimum of 11 comparisons per item for reading.

Participants

Judges were from EQAP and Fiji Mistry of education and had mix of assessment, curriculum, and education policy experience. A total of 11 judges in reading completed

on average 230 comparisons and for mathematics 13 judges did on average 290 comparisons.

Procedures

The same overall procedure used in Study 2 were used with the exception that all training was done remotely. Judges competed the exercise in EQAP premises over period of two days assisted with EQAP staff on the ground.

Results and discussion

A high correlation between the original PAT scale location and the PCM scale location was again observed in Study 2; r=0.86 and r=0.85 for reading and mathematics respectively. The correlation between PCM scale location and the original PILNA scale location was somewhat lower r=0.64 and r=0.76 for reading and mathematics respectively.

It is possible that the decrease in item in Study 3 relative to that of Study 1 and 2 might have had more of an impact on stability of PILNA items PCM location than that compared to that of PAT items. To investigate this assumption judgments across three studies were combined into single matrix owing to the fact that PAT items were used in all three studies and the PILNA items were used in study 2 and 3 thus providing strong data linkage across the studies

The joint calibration showed high correlation between PCM sale location and the original PILNA scale location r=0.85 and r=0.77 for mathematics and reading respectively. In the combined matrix the number of times PILNA items were included on average in 57 pairs in reading and 65 mathematics. The increase in such item exposure led to increase in reliability outcomes.

The increase of exposure rate with average 86 and 100 pairings per item for reading and mathematics respectively led to marginal increase of overall very high correlation between original PAT scale location and the PCM scale location was again observed in Study 3; r=0.96 for reading and r=0.94 for mathematics. These results indicate that increasing the item exposure rate might not improve outcomes past approximately 60 pairings.

PCM Operational deployment indication

The results of the research studies have provided strong evidence for the validity, reliability and operational feasibility of the Pairwise Comparison Method. These showed that CJ technique provides measurement scale outcomes that meet requirements of standard IRT linking. The PCM provides a robust solution of linking assessments with the Learning Progression Scales using the standard IRT statistical linking processes.

The Learning Progression Scales constructed in this way provided a robust statistical ordering of MPL items, but not yet the information about the location of the cut scores for the relevant MPLs. Consequently, ACER conducted a formal standard setting study (the

International Standard Setting Exercise, ISSE) to establish the location of these cut score on the Learning Progression Scales (ACER, 2022). The purpose of the ISSE and research presented here was thus to provide an independent and criterion-referenced location of SDG Indicators 4.1.1 on the Learning Progression Scales for reading and mathematics.

The Learning Progression Scales used in the ISSE were constructed using joint calibration off all items and judgments collected across the three PCM studies, as well as qualitative analyses of the ordering of all MPL items included in the study. For ease of communication, the Learning Progression Scales were transformed in the measurement scale, to a mean of 120 and standard deviation of 10 scale score points.

The research has also shown that, for operational deployment, the item exposure rates should be at set at least 40 and that the number of judges should be at least 15. The number of pairs allocated to judges seems to have had little impact on the reliability of the PAT and PILNA item PCM scale location.

References

The Australian Council for Educational Research (ACER). (2022). International Standard Setting Exercise. https://doi.org/10.37517/978-1-74286-688-8

Lazendic, G. (2019). Options for Reporting against 4.1.1 when using national assessment programs. *Global Alliance to Measure Learning (GAML): sixth meeting*.

Montoya, S. (2022). Reporting learning outcomes in basic education: country's options for indicator 4.1.1. *38th Annual Conference for Educational Assessment*.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review 34(4), 273-286.

Appendix A

The purpose of Appendix A is to provide a summary of the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959) parameters, including model fit, reliability and separation index for three studies described in the Pairwise Comparison Technical Note. Appendix A also provides graphical representation of the BTL models' fit indices (information-weighted fit–infit) for items and judges. These fit indices are functionally equivalent to item infit reported in standard IRT analyses.

Study 1: Evaluating the core premise of PCM statistical linking approach

Table 1: Study BTL model summary for reading and mathematics

Domain	Parameter	Mathematics	Reading
Mathematics	Log-likelihood	-729.54	-623.95
Mathematics	Iterations	45	86
Mathematics	Items	141	133
Mathematics	No comparisons	2994	2780
Mathematics	MLE Rel	0.96	0.97
Mathematics	Separation index	5.28	6.23

Figure 1 shows the infit density distribution across judges and items in mathematics and reading BTL models. The vertical lines denote two standard deviation cut sores used to identify the outliers in terms of the judge's severity and items' discrimination.

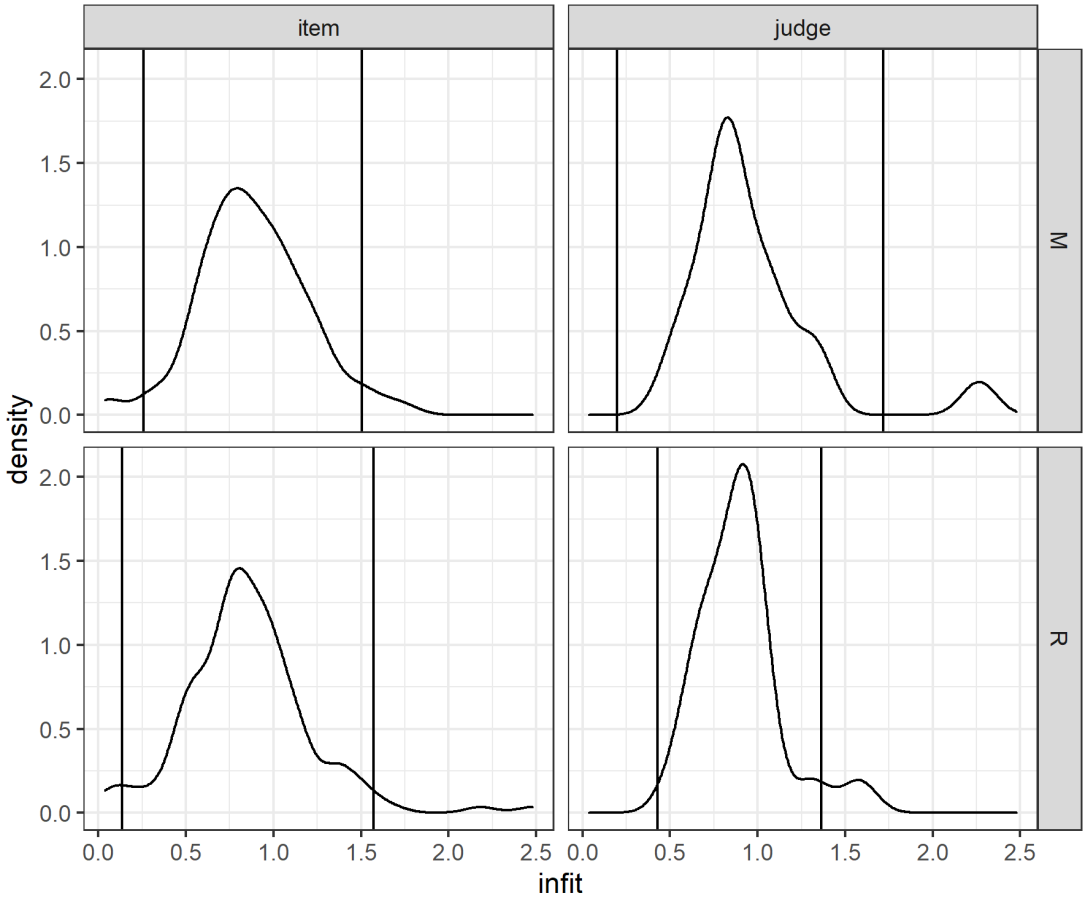

Figure 1 Study 1 infit distribution for reading and mathematics

Study 2: PCM replication and robustness

Table 2: Study 2 BTL model summary for reading and mathematics

Domain	Parameter	Mathematics	Reading
Mathematics	Log-likelihood	-1587.56	-1517.59
Mathematics	Iterations	62	53
Mathematics	Items	293	245
Mathematics	No comparisons	5250	4334
Mathematics	MLE Rel	0.96	0.96
Mathematics	Separation index	4.86	5.14

Figure 2 shows the infit density distribution across judges and items in mathematics and reading BTL models. The vertical lines denote two standard deviation cut sores used to identify the outliers in terms of the judge's severity and items' discrimination relative to that of allotter judges and items.

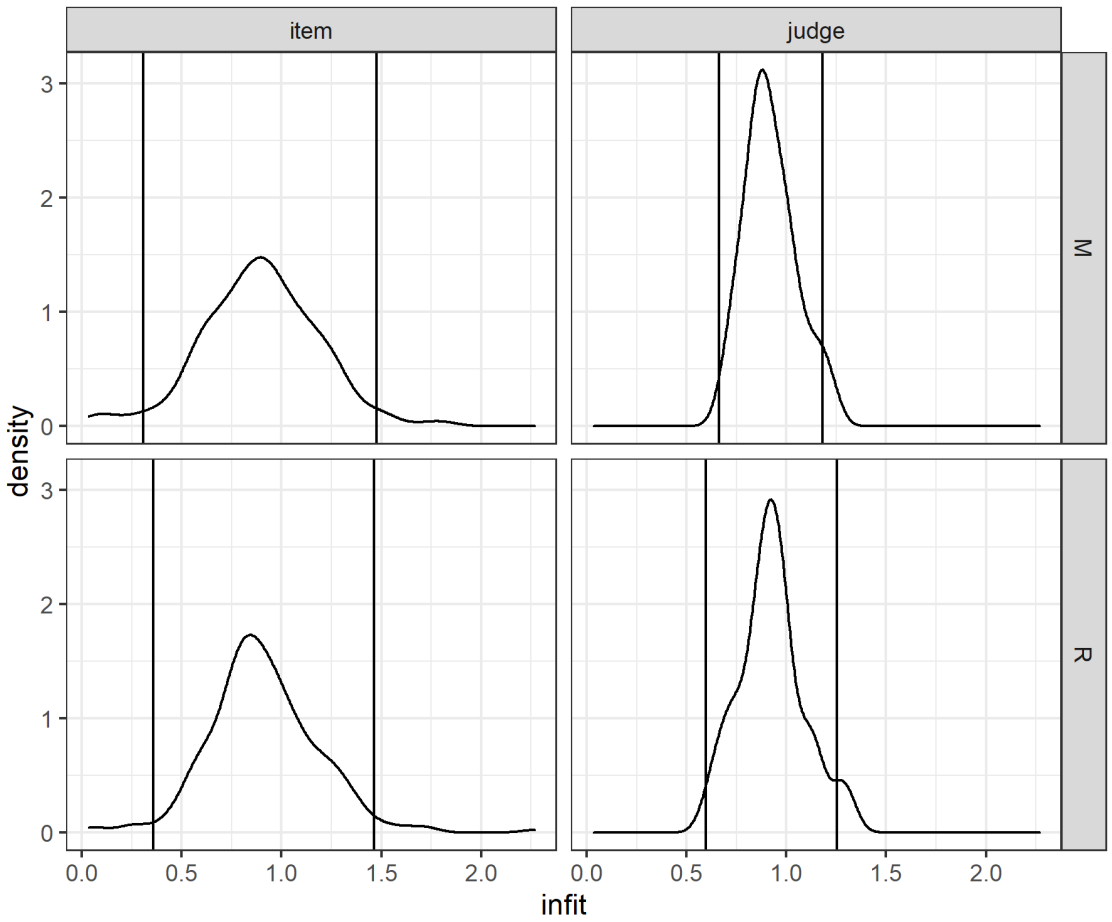

Figure 2: Study 2 infit distribution for reading and mathematics

Study 3: Operational deployment feasibility

Table 3: Study 3 BTL model summary for reading and mathematics

Domain	Parameter	Mathematics	Reading
Mathematics	Log-likelihood	-1674.26	-1160.25
Mathematics	Iterations	59	100
Mathematics	Items	279	248
Mathematics	No comparisons	3877	2400
Mathematics	MLE Rel	0.89	0.88
Mathematics	Separation index	3.09	2.93

Figure 3 shows the infit density distribution across judges and items in mathematics and reading BTL models. The vertical lines denote two standard deviation cut sores used to identify the outliers in terms of the judge's severity and items' discrimination relative to that of allotter judges and items.

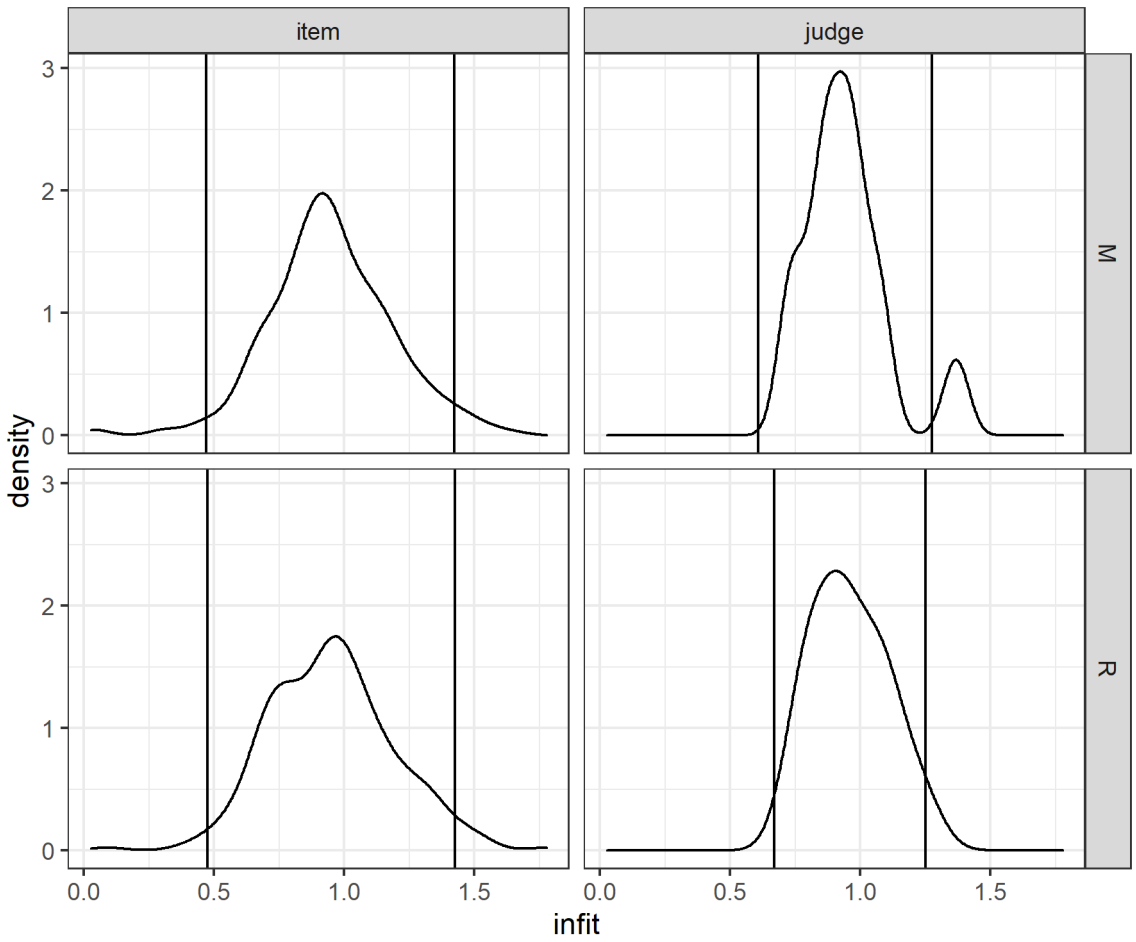

Figure 3: ***Study 3 infit distribution for reading and mathematics***

Table 4 shows the model summary for the joint calibration of the combined data matrix used to estimate the final learning progression scale.

Table 4: Joint studies response matrix BTL model summary for reading and mathematics

Domain	Parameter	Mathematics	Reading
Mathematics	Log-likelihood	-4399.07	-3559.76
Mathematics	Iterations	43	63
Mathematics	Items	406	394
Mathematics	No comparisons	12121	9514
Mathematics	MLE Rel	0.96	0.96
Mathematics	Separation index	5.25	5.31