Validating the Minimum Proficiency Levels

To enable robust and valid reporting of student achievement against the Minimum Proficiency Levels (MPLs) requirements, a systematic approach is required to establish and validate cut scores that correspond to the end of lower primary MPLa requirements, for each of reading and mathematics.

The MPLa, 'b' and 'c' cut scores for reading and mathematics were established on the Learning Progressions Scale (LPS) with an international standard setting exercise undertaken in 2022. The MPLb cut score was also established on the AMPL scale in 2021 using data collected in the MILO project. For this project, the location of the MPLa and 'b' cut scores on the AMPL scale was first established using the MPLb cut score from MILO and the difference between MPLb and MPLa in logits as established in the international standard setting exercise. A pairwise comparison method (PCM) study was then used to check and validate the location of those cut scores.

The pairwise comparison exercise

The PCM exercise was undertaken through the following four stages.

PCM training (Workshop 1): The training provided participants with an understanding of the concept of a learning progression and shared the relevant LPS for the reading and mathematics domains. The PCM activities were explained to participants, including a demonstration of the online system (ACER Signum) for making judgements. The training was conducted via an online workshop.

Undertaking the comparative judgements: After the training, participants were required to independently complete their judgements within 48 hours. The online system presented judges with pairs of items made up of an item from the assessment and/or an item already aligned to the LPS. For each pair of items, participants were required to answer the question – 'Which of the two items presented is more difficult?'

Analysing the outcomes: The judgements provided by all the participants formed a dataset that technical experts from ACER analysed to locate AMPL items on the LPS scale, providing validation of the cut-points.

Outcomes session (Workshop 2): Participants were reconvened in a plenary session to share their experiences of the process and provide feedback.

Analysis

Items and Judges

The pairwise comparison method was used to equate the LPS with the AMPL scale for both reading and mathematics. The pairwise exercise comprised 72 items from the international standard setting LPS and 41 AMPL items estimated on the AMPL scale.

For maths, there were 75 items with estimated difficulty on the LPS, and 36 items estimated on the AMPL scale. Fifteen judges participated in maths and eighteen in reading.

Pairwise Scale

Judgements on the items were completed in ACER Signum and the final datasets were downloaded and cleaned. The pairwise analysis was undertaken using ACER ConQuest using the Bradely-Terry-Luce Model. The estimates for each item included in the model for each of reading and mathematics were estimated difficulties in logits determined on the pairwise scale for each domain. The fit of both items and judges were assessed, those showing misfit were removed and the analysis was run for a second time without those items and judges to establish final locations on the pairwise scale for the remaining items.

Equating

With the MPLa and b cut scores located on the LPS and the cut score required on the AMPL Scale it was necessary to equate both the LPS and AMPL Scale with the Pairwise Scale which contained items common to each. To equate the LPS with the Pairwise Scale the item estimates for the common items were compared and outliers were detected using a Robust-Z approach - that is to identify items with significantly different scale locations between the two scales. Items showing DIF were removed, and the procedure was repeated until no items with DIF remained.

Next, the mean and standard deviation of the LP item estimates and the means and standard deviation of the Pairwise item estimates are compared, and the differences are calculated. The differences in the mean difficulties represent the shift required to move an individual estimate from one scale to the other, and a transformation can be applied. At this stage, the cut scores for mathematics and reading on the LPS were equated to the Pairwise Scale using this mean shift calculation.

Finally, it was necessary to equate the Pairwise Scale with the AMPL Scale to find the location of those cut scores for the AMPL. The same procedure was used as for the LP/Pairwise equating and the MPLa and MPLb locations on the Pairwise Scale were equated to the AMPL Scale.

Results of PCM

As can be observed in Table 21 the location of the MPLa and MPLb cut scores on the AMPL scale is no more than 0.1 logits different, when comparing the locations derived from the PCM and the locations established from a combination of the data from the MILO study and the international standard setting exercise. This close result provides evidence that the cut-scores on the AMPL scale are a valid and reliable indicator of the proficiency level required in reading and maths to meet the MPL benchmarks.

Table 1: MPL locations on the LPS and AMPL Scale (in Logits) by Method

Domain	MPL	LPS cut scores	AMPL cut scores (Common item equating from AMPL to MILO)	AMPL cut scores (PCM Equating)
Mathematics	MPLa	-3.3	-1.76	-1.74
Mathematics	MPLb	-1.6	-0.06	-0.04
Reading	MPLa	-4.7	-0.78	-0.89
Reading	MPLb	-3.0	0.92	0.81

Engagement and feedback of PCM panellists

To undertake the PCM, subject matter experts and practitioners were required from each participating country. For subject matter experts, it is preferable that participants have experience in national or standardised assessment development, curriculum development, or teacher training. Expert practitioners were also invited, particularly those with teaching experience in primary schools.

AMPLab countries were asked to engage the participation of five participants for each of mathematics and reading. The aim was to obtain a minimum of 15 participants across counties for each domain of mathematics and reading. This was achieved, with 17 panellists in mathematics and 15 panellists for reading.

The AMPLab implementing countries were responsible for paying their panellists, which was negotiated with the UIS.

After the Pairwise exercise, participants were invited back to a one hour session. In that session, ACER reported back the level of engagement in the Exercise and outlined the next steps in regard to analysis and reporting. Furthermore, it was an opportunity for participants to share their experiences with each other about undertaking the exercise , and provide feedback to ACER. This was also supplemented by an online survey, sent to participants prior to this session.

Respondents who participated in the pairwise comparison exercise were overwhelming positive about the preparation, training, support and conduct of the exercise. Every person who completed the survey, remarked that they would participate again in a similar exercise. Furthermore, many participants appreciated the opportunity to learn about and be part of an activity related to supporting measuring learning achievement. For example, one participant wrote:

> *I am really grateful for the opportunity to be part of this exercise and to make contributions to the improvement of the state of literacy and numeracy … in addressing Sustainable Development Goal 4 that aims to provide young people with quality education.*

The rating that participants gave in response to statements about the Exercise can be seen in Figure 13.
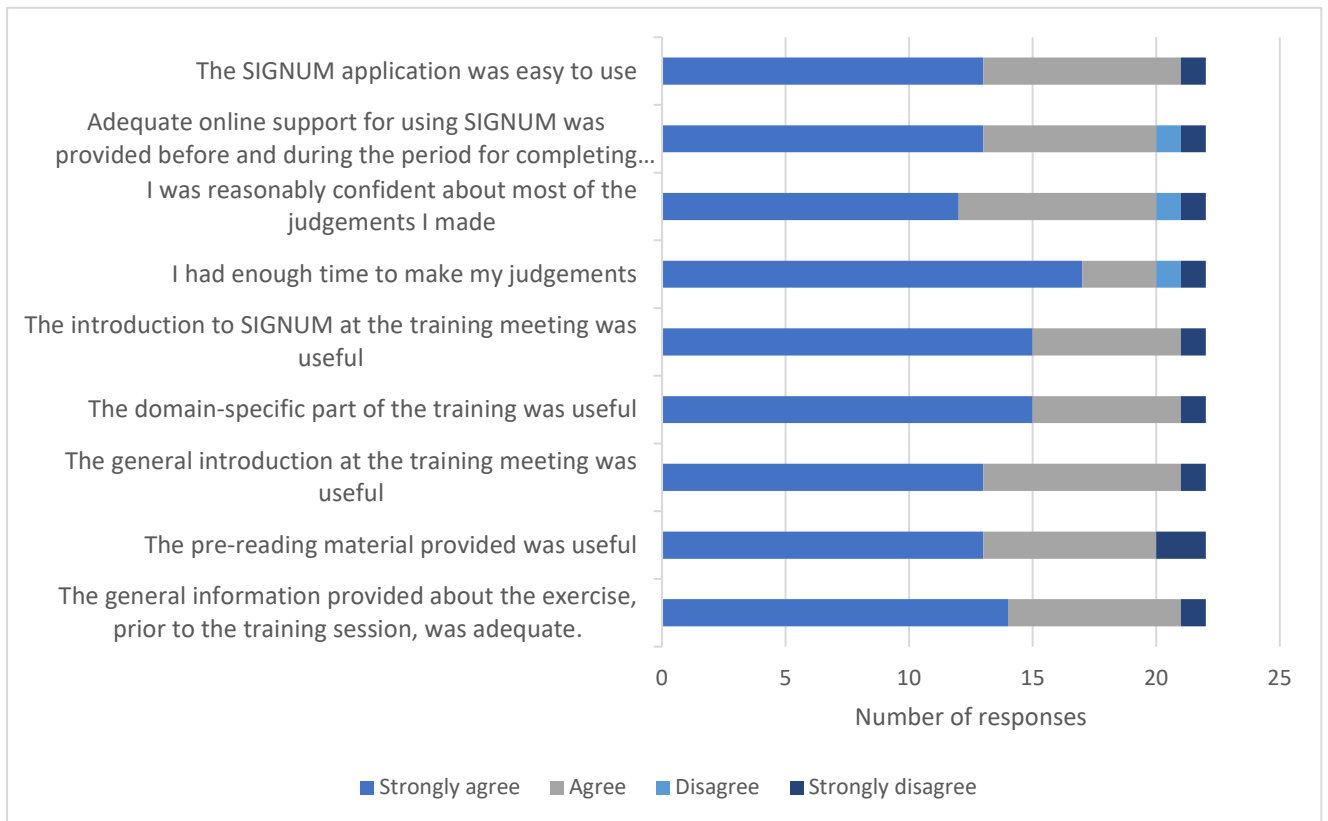

Figure 1: Panelist ratings of elements of the Pairwise Comparison Exercise