© UNICEF/UNI394567/DEJONGH

APPENDIX A

The standard setting process used to determine the MPL cut-points

INTRODUCTION

To enable robust and valid reporting of student achievement against the Minimum Proficiency Levels (MPLs) for SDG 4.1.1b, a standard setting exercise was undertaken. The standard setting exercise established cut-scores that corresponded to the end of primary MPLs for reading and mathematics. An overview of the standard setting design, method and results is provided.

STANDARD SETTING DESIGN

A modified Yes/No Angoff method (Angoff, 1971; Impara & Plake, 1997) was used to determine a single MPL cut-score for mathematics and a single MPL cut-score for reading for each AMPL. The Angoff method is based on the concept of the borderline or minimally competent student–target student.

Competence and the target student

The minimally competent student can be conceptualised as the student possessing the minimum level of knowledge and skills necessary to perform at a level 'on the borderline' between performance at the MPL and below the MPL. The borderline, target student thus belongs to the group of students that just meet the MPL requirements.

Rating the AMPL items

The Yes/No Angoff method requires participants to independently decide whether the target student is likely to answer a test item correctly. The response probability (RP) is the probability of a person of a certain ability level to respond correctly. In a standard setting exercise the RP is commonly set at 0.67 and this was the RP used in the MILO standard setting exercise.

Determining the final cut-scores

AMPL cut-scores were determined through rigorous implementation of the standard setting exercise and were then finalised following an educational impact review involving international educational community stakeholders involved in SDG 4.1.1 reporting activities who were invited to participate by the UIS.

Implementation approach

Owing to the travel restrictions caused by the pandemic, all standard setting activities were conducted as remote online sessions.

METHOD

Participants

The national project managers from each of the MILO countries nominated reading and mathematics subject matter experts and expert practitioners with experience teaching at the end of primary to participate in the training and the judgement sessions in reading and mathematics.

The breakdown of the participants across domain and language is provided in Table A.1.

TABLE A.1 Number of participants across AMPL domain and language

Domain	Language	Number of participants
Reading	English	10
Reading	French	6
Mathematics	English	7
Mathematics	French	8

Materials

During the training phases, participants had access to the original AMPL tests. During the judgment and consensus sessions the participants had access to digital versions of each item, through ACER's online standard setting system. The online system also provided information about the item keys and reading items were displayed with the relevant stimulus material.

MPL descriptions were developed independently from the AMPL, and therefore, the standard setting participants were provided with training in the MPLs and also had access to the end of primary MPL unpacking paper (ACER-GEM, 2019).

Design

The standard setting exercise consisted of training, individual judgment and consensus building sessions. Following the setting of draft cut-points in the above exercises, a standard setting impact review session was conducted. A summary of each of these steps is provided in Table A.2.

TABLE A.2 Standard setting steps and participants

Step	Summary	Participants
Training session	The standard setting participants were trained on the standard setting method and the online system used to conduct the standard setting activities.	Participants nominated from the MILO participating countries
Judgement session	Participants worked individually to analyse the AMPL items and rate each item in relation to performance by the target student.	Participants nominated from the MILO participating countries
Consensus session	Each language by domain group convened for a virtual session to attempt to find consensus on the cut-point. These sessions were facilitated by ACER and participants could update and change their responses in the online system during the consensus session.	Participants nominated from the MILO participating countries
Impact review	The outcomes of the consensus group sessions were analysed. The percentage of students at and above the MPL were calculated using the AMPL preliminary raw data (number of correct responses in a test). The standard setting method and procedure was described and outcomes of individual and consensus sessions were presented. The provisional AMPL impact data was then shared. The procedure and draft cut-scores were endorsed by the participants.	MILO country representatives and international educational community stakeholders involved in SDG 4.1.1 reporting activities invited by the UIS
Cut scores finalised	The cut scores were finalised	ACER presented the final cut scores to the UIS

© UNICEF/UN163990/GORDON

RESULTS

The participants' judgments were extracted from the online system and analysed for completeness of responses. The data for one participant in the reading group were incomplete and these data were removed from the subsequent analyses. There was no systematic difference in cut-score placement between the two language-based groups of participants for reading or for mathematics. Therefore, judgements from the two language groups were merged and all subsequent analyses used these combined data.

The summary statistics for the draft proposed cut-scores were calculated after the consensus sessions for the two domains. In order to determine the confidence interval for median and mean statistics, a non-parametric Monte Carlo bootstrap procedure was implemented to extract the lower and upper boundaries of the 95% confidence interval.

Table A.3 provides the 95% confidence interval boundaries, rounded to the nearest integer, for the median cut-scores for the two domains.

Table A.4 provides lower and upper boundaries for the mean cut-scores in reading and mathematics.

The 95% confidence interval around the mean cut-score for reading was relatively similar to that around the median. The width of the confidence interval around the mean cut-score in mathematics was smaller relative to that around the median. These outcomes indicated that using the mean cut-score statistics would provide a more stable option for calculating the position of the final cut-score. Using the mean cut-score was supported by participants during the impact review session. The mean provides a solution that uses maximum available information from the judgment sessions and solution that is in line with other international assessment reporting.

TABLE A.3 Cut-score confidence intervals: Median

Domain	N	Median	95% CI lower	95% CI upper
Reading	15	21	19	22
Mathematics	15	14	10	17

TABLE A.4 Cut-score confidence intervals: Mean

Domain	N	Mean	95% CI lower	95% CI upper
Reading	15	22	20.4	23.0
Mathematics	15	16	13.4	18.1

THE FINAL CUT-SCORES

The psychometric analyses of the complete AMPL data set found that one mathematics item and two reading items functioned differently across the two languages used in the AMPL. In order to enable the direct translation of the proposed standards' cut-scores, the decision was thus made to remove judgements for these three items from the standard setting data set. Table A.5 provides a summary of the cut-scores after removing the three items with poor psychometric properties, including the 95% confidence interval, rounded to the nearest integer.

Upon further inspection of the final impact of the proposed cut scores using the complete and weighted AMPL data, the decision was made to use the lower boundary of the 95% confidence interval for the final reading cut scores. Thus, the final reading cut score was set at 20 score points (see Table A.6). The cut points were applied to the AMPL scales and are shown in Table A.6 (for further information see Appendix B). The final cut-score statistics were used to calculate the proportions of students at and above the MPL standards.

TABLE A.5 Cut-score confidence intervals after item deletion: Mean

Domain	N	Mean	95% CI lower	95% CI upper
Reading	15	21	20	23
Mathematics	15	15	13	18

TABLE A.6 Final MPL cut-scores

Domain	Cut-score	AMPL scale score
Reading	20	0.91528
Mathematics	15	-0.06137

Endnotes

1 The proportion of children and young learners ... at the end of primary ... achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex (United Nations, 2015).

2 In 2016 for Zambia

3 Contextual data from the historical population for Zambia was not available in a format suitable for direct comparisons of populations. Some contextual data was not available from the Kenyan historical assessment.

4 The GPF advisory group on alignment was a working group comprised of psychometricians and subject matter experts who contributed to the development of the Global Proficiency Framework in 2020. The group was convened to formulate a set of alignment criteria to allow assessments to be compared to the GPF in order to determine their suitability for evaluating and reporting against SDG 4.1.1. The alignment criteria are outlined in detail in: USAID, UIS, UK Aid et al. (2020) *Policy Linking Toolkit for Measuring Global Learning Outcomes – Linking assessments to the Global Proficiency Framework.*

5 From SDG 4.1.1 Review Panel: March 2021.

6 These items were reproduced with permission from CONFEMEN.

7 For the purposes of AMPL, this item was classified as "Retrieve information" rather than "Decoding" as consistent with the GPF for reading (USAID et al, 2020a) which lists matching a given word to an illustration as an example of retrieving information.

8 The four French-speaking countries were Burkina Faso, Burundi, Côte D'Ivoire and Senegal.

9 These items are used with permission from CONFEMEN.

10 Zambia's historical assessment was conducted in 2016. All other countries' historical assessments were conducted in 2019.

11 Historical results are not reported for Kenya since the 2019 assessment of English in Kenya did not contain a sufficient number of reading comprehension item to align with the reading constructs within the GPF.

12 In the MILO project, students were the primary sampled unit. All results from the School Questionnaire are reported using student weights that are representative of the population. Therefore all results from school principals need to be interpreted in numbers of students.

13 There is no consensus among researchers and practitioners on which are the best indicators to operationalise SES. Typical children SES indicators are parents' occupation and education level, household income and home possessions. For a review of SES indicators used in educational research and other disciplines such as health, economics and sociology see Osses et al. (forthcoming).

14 Results for Kenya have been excluded based on data validation issues

15 The population chosen by countries to report against varied from Grade 5 to Grade 7.

16 A wealth index for Kenyan students was computed based on common items from the historical assessment and the AMPL. Comparisons for boys over time revealed higher scores on the wealth index in the 2021 population in comparison to the historical population.

17 For further information on different learning approaches and the benefits, considerations and enabling conditions, see for example Dabrowski et al. (2020).

18 For further recommendations relating to education in emergencies, see the Policy Monitoring tool developed for building resilient education systems (Tarricone et al., 2021).

19 Magnitude of item by gender interaction estimates from a facet model. See PISA 2006 Technical Report (OECD, 2009a).

20 'Not reached' items were defined as all consecutive missing values at the end of the test, except the first missing value of the missing series which was coded as 'embedded missing' i.e. coded the same as other items that were presented to the student but which did not receive a response. Omitting the 'not reached' items from the item calibration ensures the item difficulties not to be over-estimated.

21 The psychometric properties of the reading items administered in Burundi was unexpectedly inconsistent with those of the other countries. In particular, the response patterns in nearly all of the reading items was consistent with high rates of guessing and resulted in very low discrimination. It was therefore decided to exclude Burundi from the international reading item calibration. Burundi student reading proficiency estimations were subsequently based on the international calibration.

22 Expected a-posteriori/plausible value (EAP/PV) reliability (Adams, 2005).

23 A two-dimensional model with Quadrature estimation with 40 nodes was used.

24 So-called weighted likelihood estimates (WLEs) were used as ability estimates in this case (Warm, 1989).

25 Conceptual background and application of macros with examples are described in the PISA Data Analysis Manual SPSS®, 2nd edn (OECD, 2009b).