© UNICEF/UNI350799/DEJONGH

APPENDIX B

Technical descriptions of data analyses used to link with past national assessment results

The reading and mathematics items in the Assessments for Minimum Proficiency Levels (AMPL) were scaled using item response theory (IRT) scaling methodology. The Mixed Coefficients Multinomial Logit Model (MCMLM) as described by Adams et al. (1997) was used to scale the AMPL data. Psychometric analysis included item level analysis (item calibration at national and international level) and proficiency level generation.

The items were used to derive a one-dimensional AMPL proficiency scale for each of the two domains. This appendix outlines the procedures implemented to create the AMPL cognitive scale and provides a description of the associated processes of differential item functioning (DIF) analysis, item calibration, horizontal equating and the creation of plausible values (PVs).

THE SCALING MODEL

Test items were scaled with the one-parameter model (Rasch, 1960). In the case of dichotomous items, the model predicts the probability of selecting a correct response (value of one) instead of an incorrect response (value of zero), and is modelled as:

$$P_i(\theta_n) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

where $P_i(\theta_n)$ is the probability of person *n* scoring 1 on item *i*, θ_n is the estimated ability of person *n*, and δ_i is the estimated location of the difficulty of item *i* on this scale. For each item, item responses are modelled as a function of the latent trait θ_n.

For items with more than two (*k*) categories, the more general Rasch partial credit model (Masters & Wright, 1997) was applied, which takes the form of:

$$P_{x_i}(\theta_n) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (\theta_n - \delta_i + \tau_{ik})}$$

$$x_i = 0, 1, K, m_i$$

where $P_i(\theta_n)$ denotes the probability of person *n* scoring *x* on item *i*, θ_n denotes the person's ability, the item parameter δ_i gives the location of the difficulty of the item on the latent continuum, and τ_{ik} denotes an additional step parameter for each step *k* between adjacent categories.

The analysis of item characteristics and the estimation of model parameters were carried out with ACER ConQuest® Version 5 software (Adams et al., 2021).

SCALING COGNITIVE ITEMS

Preliminary item calibrations were first conducted separately by country and then by test language for each of the two domains. A series of item reviews were carried out to ensure the consistency of item parameters across countries to measure the same underlying construct (or latent trait).

The model fit of cognitive test items was assessed using a range of item statistics. The weighted mean-square statistic (infit) (MNSQ: Wu, 1997), which is a residual-based fit statistic, was used as a global indicator of item fit. Infit statistics were reviewed both for item and step parameters.

In addition to this, item characteristic curves (ICCs) were also used to review item fit. ICCs provide a graphical representation of item fit across the range of student abilities for each item.

Item-rest correlations were examined. Each item category has a point-biserial index, which is a comparison of the aggregate score between students selecting that category and all other students. For dichotomous items, such as multiple-choice items, the item-rest correlation is the same as the point-biserial index of the key. As a rule of thumb, the item-rest correlation should be higher than 0.20 (Ebel & Frisbie, 1986), suggesting the item discriminates relatively well between high and low performing students.

After examining the item and test level statistics and excluding some poor performing items, the mathematics test contained 26 items for French-based assessments and 29 items for English-based assessments. The reading assessments contained 28 items for French-based assessments and 27 items for English-based assessments.

DIFFERENTIAL ITEM FUNCTIONING

The quality of the items was also explored by assessing differential item functioning (DIF) by gender for each country and domain. DIF occurs when groups of students with the same ability have different probabilities of responding correctly to an item. For example, if a group of boys with the same average ability as a group of girls have a higher probability of success for a particular item, that item shows DIF in favour of boys. This constitutes a violation of the model, which assumes that the probability is only a function of ability (and item difficulty) and not of any other variable. Substantial item DIF (e.g. < -0.3 or > 0.3)[19] with respect to gender may result in bias of performance estimates across gender groups.

The gender DIF estimates range between -0.084 and 0.246 for AMPL Mathematics and between -0.063 and 0.104 for AMPL Reading. No instances of substantial gender DIF were encountered so no items were removed for this reason.

ITEM CALIBRATION

Missing student responses, likely caused by issues with test length ('not reached' items),[20] were omitted from the calibration of item parameters but were treated as incorrect for the scaling of student responses. All other missing responses were included as incorrect responses for the calibration of items (except for the ones that were not administered).

Item parameters were calibrated using all countries' sampled data of students identified as respondents,[21] taking student grades into account. Student grade dummies were created to reflect different target student populations across the MILO participating countries, ranging from Grades 5 to 7. The student sample weights were rescaled so that each country had the same sum of weights to ensure that each country was equally represented in the sample (senate weighting). The items were calibrated separately for each domain with the item mean set to zero. After removing items with unsatisfactory scaling characteristics, a total of 29 Mathematics items and 29 Reading items were used across both languages for international scaling.

Table B.1 and Table B.2 show the item thresholds on the AMPL scales with a response probability of 0.50 in logits. For example, a student with an ability estimate equal to the difficulty estimate of an item would have a 0.5 probability of answering the item correctly. It also shows the respective percentages of correct responses (facility) for domain sample (giving equal weight to each country). The item-rest correlation, the weighted fit statistics and the flag for gender DIF are included in the last three columns.

© UNICEF/UNI308001/SCHERMBRUCKER

TABLE B.1 Item thresholds in logits – Assessments for Minimum Proficiency Levels (AMPL) Reading (excluding Burundi)

Item	Max Score	Threshold 1	Facility*	Item-rest correlation	Weighted Fit (MNSQ)	Gender DIF
R_MR001	1	-1.45	72%	0.52	0.83	No
R_MR002	1	-1.18	68%	0.56	0.80	No
R_MR003	1	-1.10	67%	0.55	0.83	No
R_MR024	1	0.99	28%	0.49	0.91	No
R_MR025	1	-0.39	53%	0.46	0.96	No
R_MR033	1	0.58	32%	0.24	1.12	No
R_MR034	1	1.36	26%	0.34	1.12	No
R_MR035	1	0.57	35%	0.42	1.01	No
R_MR041	1	0.83	30%	0.34	1.08	No
R_MR042	1	1.07	26%	0.35	1.05	No
R_MR043	1	-1.19	68%	0.42	0.96	No
R_MR044	1	-0.64	58%	0.35	1.08	No
R_MR056	1	-1.35	71%	0.55	0.80	No
R_MR058	1	-0.54	56%	0.59	0.82	No
R_MR059	1	-0.16	49%	0.38	1.06	No
R_MR069	1	-0.34	52%	0.42	1.00	No
R_MR087	1	0.65	34%	0.48	0.93	No
R_MR089	1	0.59	35%	0.51	0.92	No
R_MR090	1	0.16	42%	0.45	0.99	No
R_MR201	1	0.02	45%	0.38	1.06	No
R_MR202	1	0.58	35%	0.44	0.98	No
R_MR203	1	-0.10	47%	0.32	1.13	No
R_MR204	1	0.83	30%	0.21	1.22	No
R_PF449	1	-1.60	75%	0.36	1.00	No
R_PF455	1	0.38	38%	0.53	0.90	No
R_PF456	1	0.69	32%	0.32	1.10	No
R_PF458	1	1.13	23%	0.29	1.03	No
R_PF487	1	-0.76	60%	0.35	1.08	No
R_PF489	1	0.35	39%	0.37	1.07	No

*Note: Facility, percentages of correct responses, was computed with countries equally weighted.

TEST RELIABILITY

The ConQuest® separation reliability estimate[22] of the test, as obtained from the scaling model, was approximately between 0.83 and 0.86 for AMPL Reading and AMPL Mathematics. Separation reliability values above 0.8 are considered to indicate appropriate reliability.

TABLE B.2 Item thresholds in logits – Assessments for Minimum Proficiency Levels (AMPL) Mathematics

Item	Max Score	Threshold 1	Threshold 2	Facility*	Item-rest correlation	Weighted Fit (MNSQ)	Gender DIF
M_MM004	1	-1.89		74%	0.45	0.88	No
M_MM011	1	-0.33		43%	0.41	0.95	No
M_MM016	1	-0.05		38%	0.34	1.02	No
M_MM019	1	0.22		33%	0.44	0.93	No
M_MM022	1	-0.63		50%	0.38	1.00	No
M_MM029	1	0.09		35%	0.23	1.11	No
M_MM030	2	-0.22	0.85	34%	0.34	1.48	No
M_MM060	1	-1.10		59%	0.38	1.00	No
M_MM075	1	1.06		19%	0.14	1.11	No
M_MM089	1	0.28		31%	0.20	1.15	No
M_MM090	1	0.82		22%	0.34	0.99	No
M_MM101	1	0.59		31%	0.31	1.13	No
M_MM104	1	0.61		31%	0.33	1.11	No
M_MM125	1	-0.85		53%	0.48	0.90	No
M_MM175	1	1.19		18%	0.24	1.06	No
M_MM191	1	1.17		18%	0.28	1.03	No
M_MM197	1	1.21		17%	0.27	1.03	No
M_MM206	1	2.48		6%	0.25	0.97	No
M_MM208	1	-0.74		52%	0.50	0.89	No
M_MM209	2	-1.06	1.10	37%	0.25	1.26	No
M_PM422	1	-1.59		69%	0.49	0.87	No
M_PM445	1	-0.44		45%	0.51	0.87	No
M_PM449	1	0.30		31%	0.26	1.10	No
M_PM454	1	0.41		29%	0.44	0.92	No
M_PM459	1	-0.83		54%	0.49	0.89	No
M_PM462	1	-1.40		65%	0.50	0.86	No
M_PM468	1	-0.57		48%	0.49	0.91	No
M_PM469	1	-0.08		38%	0.50	0.90	No
M_PM942	1	-0.28		43%	0.36	1.00	No

*Note: Facility, percentages of correct responses, was computed with countries equally weighted.

POPULATION MODEL AND CONDITIONING

Plausible values methodology was used to generate estimates of students' Reading and Mathematics proficiency. Using item parameters anchored at their estimated values from the calibration process, a set of five plausible values were randomly drawn from the marginal posterior of the latent distribution (Mislevy, 1991; Mislevy & Sheehan, 1987; von Davier et al., 2009). Here, 'not reached' items were included as incorrect responses, just like other (embedded) missing responses. Estimations were based on the conditional item response model and the population model, which included a regression equation including background and survey variables used for conditioning (Adams & Wu, 2002). The ACER ConQuest software (Adams et al., 2021) was used to draw the plausible values.

A two-dimensional conditioning model[23] was built for each country. Some variables were used as direct regressors in the conditioning model for drawing plausible values. These included dummy variables of explicit sampling strata of country, the school mean performance variable adjusted for the student's own performance (WLE[24]), school type, school location and student gender. Most of the other student background variables such as student age and responses to questions in the Student Questionnaire were re-coded into dummy variables which were transformed into components by a principal component analysis (PCA). The principal components were estimated for each country separately. Subsequently, the components that explained 99 per cent of the variance in all the original variables were included as regressors in the conditioning model.

HORIZONTAL EQUATING

LINK data from the 2021 national or regional assessments were calibrated separately for each national country sample. The calibration outcomes were used to review item statistics and detect any problematic items. After item review, four Mathematics items for Zambia and one Mathematics item for PASEC were excluded.

The same item treatments of item exclusion were applied to calibrations on the historical data. The historical data for Zambia and Kenya were calibrated separately and student plausible values were generated. The PASEC 2019 data did not require re-calibration as the PASEC scale was already established in 2014. PASEC 2019 item parameters and student plausible values on the historic scale were available.

Using item parameters anchored at their estimated values from the calibration process on the historical data, the conditioning model was applied and generated a set of five plausible values for both 2021 LINK data by country.

To equate the 2021 PASEC LINK data to the historic PASEC scale, the following equating shift was added to the plausible values for each domain.

PASEC Mathematics = 0.075; PASEC Reading = 0.114

Equating the 2021 PASEC LINK data also required further adjustments of test correction constants as the test included a reduced set of items and had a shorter test time.

Burundi:

$PV_{PASEC\_link\_adj}=0.9158 * (PV_{PASEC\_link\_his}-0.5734) + 0.4379$

Burkina Faso:

$PV_{PASEC\_link\_adj}=0.9396 * (PV_{PASEC\_link\_his}-0.5170) + 0.4499$

Côte d'Ivoire:

$PV_{PASEC\_link\_adj}=0.9027 * (PV_{PASEC\_link\_his}+0.6808) - 0.5101$

Senegal:

$PV_{PASEC\_link\_adj}=0.9833 * (PV_{PASEC\_link\_his}-0.5204) + 0.5571$

Where $PV_{PASEC\_link\_ad}$ is the 2021 PASEC LINK PV adjusted by the test correction constants, $PV_{PASEC\_link\_his}$ is the 2021 PASEC LINK PV on historic PASEC scale.

TABLE B.3 Mean and standard deviations of Assessments for Minimum Proficiency Levels (AMPL) and LINK scales by domain

Country	MATHEMATICS				READING			
	AMPL		LINK data		AMPL		LINK data	
	MEAN (MN_{AMPL})	STANDARD DEVIATION (SD_{AMPL})	MEAN (MN_{LINK})	STANDARD DEVIATION (SD_{LINK})	MEAN (MN_{AMPL})	STANDARD DEVIATION (SD_{AMPL})	MEAN (MN_{LINK})	STANDARD DEVIATION (SD_{LINK})
Burkina Faso	-0.479	0.626	0.725	0.895	-0.154	0.817	1.237	1.057
Burundi	-0.748	0.639	0.34	0.729	-0.917	0.516	0.201	0.721
Côte d'Ivoire	-1.536	1.14	-0.435	0.811	-0.787	1.395	0.407	1.479
Kenya	0.472	0.833	-0.218	1.008	0.902	1.224		
Senegal	-0.387	0.764	0.509	0.913	-0.123	0.934	1.216	1.04
Zambia	-1.305	0.511	-0.69	0.647	-0.898	0.7	-0.669	0.828

1. $PV_{LINK\_AMPL} = ((PV_{LINK} - MN_{LINK}) / SD_{LINK}) *SD_{AMPL} + MN_{AMPL}$
2. $PV_{LINK\_AMPL} = ((PV_{PASEC\_link\_adj} - MN_{LINK}) / SD_{LINK}) *SD_{AMPL} + MN_{AMPL}$

Common person equating by country was conducted to place the 2021 LINK results on AMPL scales. The LINK PVs were adjusted for each country using the weighted means and the weighted standard deviations. The equating quality was then examined. Table B.3 provides the mean and standard deviations of the AMPL and LINK scales by domain. The values are reported in logits.

Where PV_{LINK_AMPL} is the adjusted 2021 LINK PV on the AMPL scale, PV_{LINK} is the 2021 LINK PV of Kenya or Zambia, $PV_{PASEC\_link\_adj}$ is the 2021 PASEC LINK PV described in the previous paragraph.

The same transformations were then applied to all historic plausible values by country in order to place them onto the AMPL scales.

MPL CUT-POINTS

The proficiency cuts were determined by the standard setting as described in Appendix A. The cut points below were derived from the WLE equivalence tables. They corresponded to raw scores of reading (0.91528) and mathematics (-0.06137) which were 20 and 15 items correct, respectively. The cuts were applied to both the AMPL and the adjusted historic plausible values for each domain.

SAMPLING VARIANCE AND MEASUREMENT VARIANCE

Unbiased standard errors include both sampling variance and measurement variance. The sampling variance on population estimates from cluster samples is obtained by utilising the application of replication techniques (Gonzalez & Foy, 2000; Wolter, 1985). The other component of the standard error, the measurement variance, can be derived from the variance between the five plausible values of AMPL. The sampling variances of population statistics in AMPL were estimated using the jackknife repeated replication technique (JRR). Specialist software, the SPSS® Replicates add-in, was used to run tailored SPSS® macros for statistics estimations.[25]

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).