Scaling AMPLab Items: Technical Note

November 2023

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).

Contents

Acknowledgments

This project, Development of an Assessment of Minimum Proficiency Level (AMPL) aligned to Global Proficiency Framework (GPF) Level 2, is referred to AMPLab. This UNESCO Institute for Statistics (UIS) project is supported by the Bill & Melinda Gates Foundation.

The Australian Council for Educational Research (ACER) is the technical partner for this project. Support is provided from ACER's Global Education Monitoring Centre (GEM Centre), an ACER initiative in partnership with the Australian government's Department of Foreign Affairs and Trade.

Background

As part of SDG 4, Indicator 4.1.1 aims to measure the "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 in (i) reading and (ii) mathematics, by sex." To meet this goal, UIS has coordinated efforts to establish common reading and mathematics scales for all three points of Indicator 4.1.1, building on existing cross-national and national assessments. As a result of these efforts, two important points of consensus have been reached: the definition of the Minimum Proficiency Level (MPL) and the Global Proficiency Framework (GPF).

The overarching objective of the AMPLab project is to measure and analyse the proficiency of students at the end of lower primary and end of primary. This will:

- produce baseline population estimates in reading and mathematics proficiency to enable participating countries to set informed targets for improvement
- facilitate reporting against SDG 4.1.1
- aid the tracking of learning progress over time
- complement tools that had been already developed in 2021 in the Monitoring the Impacts of COVID-19 on Learning Outcomes (MILO) study.

Introduction

1. This report outlines the technical treatment of reading and mathematics assessment data from the AMPLab study administered in 2023 in 4 countries and across 5 populations: The Gambia (grade 3), Kenya (grade 6), Lesotho (grade 7) and Zambia (grades 4 and 7). This document satisfies deliverable – *12.4 Calibration and shift or anchor equating outcomes*.

2. The reading and mathematics items in the Assessments for Minimum Proficiency Levels (AMPL) were psychometrically scaled using item response theory (IRT) methodology. The Mixed Coefficients Multinomial Logit Model (MCMLM) as described by Adams, Wilson and Wang (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.

3. The items were used to derive a one-dimensional AMPL proficiency scale for each of the two domains: reading and mathematics. This Technical Note outlines the procedures implemented to create the AMPL cognitive scales 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

4. 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:

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

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.

5. 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:

$$P_{x_i}(\theta_n) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (\theta_n - \delta_i + \tau_{ik})} \quad x_i = 0, 1, \mathrm{K}\ , m_i$$

where $P_{xi}\ (\theta_n)$ 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.

6. 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

7. Preliminary item calibrations were first conducted separately by country 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).

8. 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.

9. 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.

10. 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, due to the nature of scoring, the item-rest correlation is expected to be same as the point-biserial index of the key. For partial-credit item, the item-rest correlation is expected to be different from the point-biserial index of any score points. 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.

11. After examining the item and test level statistics and excluding some poor performing items, the mathematics test contained 58 items for end of primary assessments and 29 items for lower primary assessments. The reading assessments contained 72 items for end of primary assessments and 41 items for lower primary assessments.

Differential item functioning

12. 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[1]) with respect to gender may result in bias of performance estimates across gender groups. The gender DIF estimates range between -0.121 and 0.195 for AMPL Mathematics and between -0.184 and 0.180 for AMPL Reading. No instances of substantial gender DIF were encountered so no items were removed for this reason.

Item calibration

13. Missing student responses, likely caused by issues with test length ('not reached' items)[2], were omitted from the calibration of item parameters but were treated as incorrect for the scaling of student responses. All other missing responses to administered items were included as incorrect responses for the calibration of items .

14. Item parameters were calibrated using all countries' data from sampled students identified as respondents[3], taking lower and end of primary populations into account. Student dummy variables of lower or end of primary were created to reflect different target student populations across the AMPLab participating countries, ranging from grade 3 to 7. The student sample weights were rescaled so that each country has the same sum of weights to ensure that each country was equally represented in the sample (senate weighting). After performing the final item review and the horizontal equating study by domain, the items were calibrated separately for each domain of which common items to MILO were anchored with MILO item parameters. A total of 58 Mathematics items and 72 Reading items were used across both lower and upper primary for international scaling. Horizontal equating study is discussed in the section below.

Test reliability

15. The ConQuest® separation reliability estimate[4] of the test, as obtained from the scaling model, was around 0.9 for AMPLab Mathematics and AMPLab Reading.

[1] PISA 2006 Technical Report (OECD, 2009).

[2] '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 is to ensure the item difficulties not to be over-estimated.

[3] 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.

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

Horizontal equating

16. Test items consisted of new and previously administered items. The previously administered items were developed for and used in MILO study. As they had been kept confidential, they could be used as horizontal link items to equate the results of the AMPLab 2023 assessment with the AMPL scale established in the MILO study.

17. To ensure that the link items had the same measurement properties across two studies, the relative difficulties in AMPLab 2023 and MILO 2021 were compared for each domain. Seven out of 25 Mathematics common items and one out of 25 Reading common items showed large DIF between AMPLab 2023 and MILO 2021 and were unlinked for equating. The link item set of each domain has slightly lower discrimination in AMPLab 2023 than MILO 2021. The average discrimination (item–rest correlation) were 0.38 in MILO 2021 and 0.38 in AMPLab 2023 for Mathematics, and 0.43 in MILO 2021 and 0.39 in AMPLab 2023 for Reading. The average DIF with respect to gender in both studies was close to zero in both domains.

18. The final item calibration was done separately by domain. The item parameters of the finalised horizontal link item set were anchored at their estimated values from the calibration process of the MILO study to establish the AMPL scale. The remaining item estimates were obtained on the same AMPL scale at the international level.

19. Table 1 and Table 2 shows the item thresholds on the AMPL scales with a response probability of 0.50 in logits. For example, a student whose ability estimate is equal to the item difficulty estimate of an item, the student would have 50% chance to answer that 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.

Table 1: Item thresholds in logit – AMPLab Mathematics

Item	Strand	Horizontal Link Item	Max Score	Threshold 1	Threshold 2	Facility*	Item-rest correlation	Weighted Fit (MNSQ)	Gender DIF
AM001	Number	No	1	-3.22		77%	0.52	0.81	No
AM002	Number	No	1	-4.11		87%	0.39	0.89	No
AM003	Number	No	1	-2.27		56%	0.56	0.88	No
AM004	Number	No	1	-1.78		47%	0.56	0.90	No
AM005	Number	No	1	-1.60		49%	0.38	1.12	No
AM006	Number	No	1	-3.43		76%	0.45	0.93	No
AM007	Number	No	1	-2.33		57%	0.47	1.01	No
AM008	Number	No	1	-1.97		50%	0.51	0.98	No
AM009	Number	No	1	-3.56		77%	0.48	0.87	No
AM010	Number	No	1	-3.27		77%	0.46	0.87	No
AM011	Number	No	1	-2.97		73%	0.47	0.89	No
AM012	Number	No	1	-3.08		70%	0.57	0.81	No
AM013	Number	No	1	-1.77		47%	0.57	0.88	No
AM014	Number	No	1	-1.82		48%	0.57	0.91	No
AM015	Number	No	1	0.39		31%	0.20	1.13	No
AM016	Number	No	1	-1.35		45%	0.50	0.95	No
AM017	Measure	No	1	-2.26		61%	0.39	1.12	No
AM018	Measure	No	1	-1.02		34%	0.40	1.08	No
AM019	Measure	No	1	-2.54		61%	0.41	1.08	No
AM020	Measure	No	1	-2.11		58%	0.50	0.91	No
AM021	Geo	No	1	-1.40		46%	0.22	1.40	No
AM022	Geo	No	1	-3.02		69%	0.48	0.93	No
AM023	Geo	No	1	-1.38		45%	0.45	1.00	No
AM024	Geo	No	1	-1.13		41%	0.46	0.98	No
AM025	Stat	No	1	-1.76		47%	0.47	1.02	No
AM026	Stat	No	1	-1.52		42%	0.42	1.09	No
AM027	Stat	No	1	-2.19		60%	0.49	0.96	No
AM028	Stat	No	1	-1.42		46%	0.52	0.90	No
AM029	Algebra	No	1	-2.01		56%	0.43	1.04	No
AM030	Algebra	No	1	-2.09		58%	0.52	0.92	No
MM004	Number	No	1	-0.76		49%	0.41	0.94	No
MM011	Algebra	No	1	-2.17		76%	0.34	0.95	No
MM016	Number	Yes	1	-0.05		36%	0.23	1.08	No
MM019	Number	Yes	1	0.22		31%	0.28	1.05	No
MM022	Number	Yes	1	-0.63		39%	0.33	0.97	No
MM029	Stat	No	1	-0.90		51%	0.06	1.23	No
MM060	Number	No	1	-0.09		35%	0.19	1.08	No
MM089	Geo	Yes	1	0.28		24%	0.06	1.08	No
MM090	Stat	Yes	1	0.82		19%	0.23	0.99	No
MM104	Algebra	Yes	1	0.61		18%	0.17	0.96	No
MM125	Stat	Yes	1	-0.85		64%	0.37	0.97	No
MM175	Number	Yes	1	1.19		21%	0.16	1.46	No
MM191	Measure	Yes	1	1.17		17%	0.18	1.19	No
MM197	Geo	No	1	0.35		27%	0.18	1.10	No
MM208	Number	Yes	1	-0.74		38%	0.32	0.97	No
MM209	Geo	Yes	2	-1.06	1.10	45%	0.14	1.20	No
MM210	Number	No	1	0.29		28%	0.30	1.01	No
MM211	Measure	No	1	0.16		30%	0.13	1.11	No
MM212	Measure	No	1	-0.60		45%	0.39	0.95	No
PM422	Number	No	1	-0.49		43%	0.43	0.92	No
PM445	Stat	Yes	1	-0.44		55%	0.49	0.96	No
PM449	Measure	Yes	1	0.30		28%	0.14	1.11	No
PM454	Number	Yes	1	0.41		26%	0.27	1.02	No
PM459	Geo	No	1	-2.07		75%	0.34	0.97	No
PM462	Measure	Yes	1	-1.40		51%	0.39	1.04	No
PM468	Number	Yes	1	-0.57		31%	0.36	0.92	No
PM469	Measure	Yes	1	-0.08		24%	0.19	0.95	No
PM942	Number	Yes	1	-0.29		42%	0.26	1.06	No

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

Table 2: Item thresholds in logit – AMPLab Reading

Item	Strand	Horizontal Link Item	Max Score	Threshold 1	Facility*	Item-rest correlation	Weighted Fit (MNSQ)	Gender DIF
ALD001_3_1	Decoding	No	1	-1.64	67%	0.40	0.98	No
ALD002_3_2	Decoding	No	1	-1.97	72%	0.45	0.89	No
ALD003_3_3	Decoding	No	1	-1.69	67%	0.50	0.85	No
ALD004_3_4	Decoding	No	1	-1.96	72%	0.38	0.95	No
ALD005_3_5	Decoding	No	1	-1.68	67%	0.42	0.92	No
ARD001	Decoding	No	1	-1.73	62%	0.53	0.86	No
ARD002	Decoding	No	1	-2.14	70%	0.50	0.86	No
ARD003	Decoding	No	1	-1.57	59%	0.44	0.97	No
ARD004	Decoding	No	1	-0.74	43%	0.54	0.87	No
ARD005	Decoding	No	1	-1.14	51%	0.50	0.92	No
ALL001_1_1	Listening	No	1	-1.04	55%	0.22	1.21	No
ALL002_1_2	Listening	No	1	-0.50	44%	0.28	1.17	No
ALL003_1_3	Listening	No	1	-1.13	57%	0.38	1.06	No
ALL004_1_4	Listening	No	1	-0.61	47%	0.19	1.26	No
ALL005_1_5	Listening	No	1	-1.07	56%	0.41	1.02	No
ALL006_2_1	Listening	No	1	-1.65	67%	0.34	1.05	No
ALL007_2_2	Listening	No	1	-1.42	63%	0.43	0.96	No
ALL009_2_4	Listening	No	1	-1.24	59%	0.38	1.04	No
ALL010_2_5	Listening	No	1	-0.13	37%	0.34	1.06	No
ARM002	Reading	No	1	1.09	24%	0.59	0.80	No
ARR001	Reading	No	1	-1.43	60%	0.46	0.95	No
ARR002	Reading	No	1	-3.08	86%	0.36	0.93	No
ARR003	Reading	No	1	-1.78	67%	0.51	0.84	No
ARR004	Reading	No	1	-1.99	71%	0.50	0.85	No
ARR005	Reading	No	1	-2.33	73%	0.44	0.91	No
ARR006	Reading	No	1	-1.15	55%	0.41	1.00	No
ARR007	Reading	No	1	-1.16	51%	0.54	0.87	No
ARR008	Reading	No	1	-2.52	79%	0.42	0.89	No
ARR009	Reading	No	1	-1.25	53%	0.40	1.03	No
ARR010	Reading	No	1	-0.89	50%	0.47	0.93	No
ARR011	Reading	No	1	-0.94	51%	0.50	0.91	No
ARR012	Reading	No	1	-1.37	59%	0.47	0.92	No
ARR013	Reading	No	1	-0.51	43%	0.44	0.97	No
ARR014	Reading	No	1	-1.40	60%	0.43	0.97	No
ARR015	Reading	No	1	-1.12	50%	0.41	1.02	No
ARR016	Reading	No	1	-1.28	53%	0.34	1.08	No
ARR017	Reading	No	1	-1.06	49%	0.42	1.00	No
ARR019	Reading	No	1	0.02	29%	0.41	0.97	No
ARR020	Reading	No	1	-1.19	52%	0.50	0.91	No
ARR021	Reading	No	1	0.18	41%	0.14	1.27	No
ARR022	Reading	No	1	-0.51	43%	0.28	1.17	No
ARR023	Reading	No	1	0.04	44%	0.25	1.13	No
ARR024	Reading	No	1	-0.39	36%	0.20	1.26	No
ARR025	Reading	No	1	-0.72	47%	0.28	1.17	No
MR001	Reading	Yes	1	-1.45	74%	0.48	0.84	No
MR002	Reading	Yes	1	-1.18	72%	0.51	0.79	No
MR003	Reading	Yes	1	-1.10	64%	0.53	0.86	No
MR004	Reading	No	1	-2.12	82%	0.36	0.92	No
MR005	Reading	No	1	0.23	38%	0.45	0.96	No
MR006	Reading	No	1	0.65	31%	0.34	1.06	No
MR024	Reading	Yes	1	0.99	31%	0.43	1.08	No
MR025	Reading	Yes	1	-0.39	45%	0.44	0.96	No
MR035	Reading	Yes	1	0.57	32%	0.47	0.91	No
MR041	Reading	Yes	1	0.83	32%	0.25	1.20	No
MR042	Reading	Yes	1	1.07	26%	0.34	1.10	No
MR043	Reading	Yes	1	-1.19	59%	0.46	1.05	No
MR044	Reading	Yes	1	-0.64	54%	0.28	1.10	No
MR056	Reading	Yes	1	-1.35	76%	0.49	0.76	No

Item	Strand	Horizontal Link Item	Max Score	Threshold 1	Facility*	Item-rest correlation	Weighted Fit (MNSQ)	Gender DIF
MR058	Reading	Yes	1	-0.54	56%	0.55	0.84	No
MR059	Reading	No	1	-0.80	59%	0.19	1.15	No
MR069	Reading	Yes	1	-0.34	50%	0.34	1.05	No
MR087	Reading	Yes	1	0.65	35%	0.44	0.98	No
MR089	Reading	Yes	1	0.59	26%	0.36	0.94	No
MR090	Reading	Yes	1	0.16	35%	0.45	0.89	No
MR201	Reading	Yes	1	0.02	42%	0.32	1.08	No
MR202	Reading	Yes	1	0.58	35%	0.31	1.08	No
MR204	Reading	Yes	1	0.84	32%	0.19	1.28	No
PF449	Reading	Yes	1	-1.60	78%	0.38	0.87	No
PF455	Reading	Yes	1	0.38	32%	0.47	0.87	No
PF456	Reading	Yes	1	0.69	32%	0.19	1.21	No
PF487	Reading	Yes	1	-0.76	66%	0.27	1.07	No
PF489	Reading	Yes	1	0.35	31%	0.37	0.96	No

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

Population model and conditioning

20. Plausible values (PV) 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.

21. A two-dimensional conditioning model[5] 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[6]), school type and student gender. Most of the other student background variables such as student age and responses to questions in the student questionnaire are re-coded into dummy variables which are 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.

Booklet effects

22. A total of 6 test booklets were randomly assigned to the upper primary students. Of those 6 booklets, 2 test booklets were given to the lower primary students. It was observed that the average facility of each of these 2 booklets is significantly higher

[5] A two-dimensional model with Quadrature estimation with 40 nodes was used.
[6] So called weighted likelihood estimates (WLEs) were used as ability estimates in this case (Warm, 1989).

than the other booklets at the upper primary stage within each country. It is therefore concluded that booklet effects were present in the upper primary stage. As booklet effect can have influences on the estimated proficiency distributions, it required to adjust the ability estimates of the upper primary students by booklet and country accordingly.

23. Weighted average PV mean of the upper primary students by booklet and country were compared with the corresponding weighted country mean. For each booklet, an adjustment shift was computed and applied to each PV estimate of the students who was assigned to that booklet. Table 3 lists the adjustment shifts applied to the ability estimate of the upper primary students by booklet and country.

Table 3: Adjustments of booklet effect for upper primary students

Country	Booklet	Mathematics	Reading
Kenya	3	-0.14303	-0.10416
	4	-0.12011	-0.10618
	5	-0.02698	-0.02656
	6	-0.01665	-0.03256
	7	0.14146	0.14331
	8	0.16823	0.12850
Lesotho	3	-0.05555	-0.05799
	4	-0.01928	-0.05676
	5	-0.04567	-0.00746
	6	-0.06726	-0.04044
	7	0.11291	0.14211
	8	0.07527	0.02197
Zambia	3	0.04334	-0.00977
	4	-0.01360	-0.05891
	5	-0.05705	0.00043
	6	0.03255	0.07909
	7	0.03948	0.03664
	8	-0.03542	-0.03462

MPL cut-points

24. The same MPL cut points used in MILO were applied to the proficiency cuts AMPLab standard “b” (-0.06137 for Mathematics and 0.91528 for Reading).

25. The cut points of AMPLab standard “a” (-1.76137 for Mathematics and -0.78472 for Reading) were determined by applying the same distance between the cuts of MPLa and MPLb, which is 1.7 logit, on the cut point of AMPLab standard “b” (ACER, 2022).

Sampling variance and measurement variance

26. 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[7].

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

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Measurement, Evaluation, and Statistical Analysis*, *31*(2), 162–172.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Adams, R. J., & Wu, M. L. (2002). *PISA 2000 Technical Report*. OECD.

Adams, R. J., Wu, M. L., Macaskill, G., Haldane, S. A., Sun, X. X., & Cloney, D. (2021). *ACER ConQuest Version 5: Generalised item response modelling software* (Version 5) [Computer software]. ACER.

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

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of Education Measurement*. Prentice Hall.

Gonzalez, E. J., & Foy, P. (2000). Estimation of sampling variance. In M. O. Martin, K. D. Gregory, & S. E. Semler (Eds.), *TIMSS 1999 Technical Report*. Chestnut Hill.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 101–122). Springer.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *The NAEP 1983-1984 Technical Report*. Educational Testing Service.

OECD. (2009). *PISA Data Analysis Manual SPSS® (2nd edn)*. OECD.

OECD. (2009). *PISA 2006 Technical Report*. OECD.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Nielsen and Lydiche.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). *What are plausible values and why are they useful?* (Vol. 2). IER Institute and ETS.

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrika*, *54*, 427–450.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag.