

WG/GAML/11/2.3

TECHNICAL DOCUMENTATION SUPPORTING THE EXEMPLAR FOUNDATIONAL MATHEMATICS ITEMS



February 15, 2025



1. Overview¹

1.1 Background

During the May 2024 benchmarking meeting of the Technical Advisory Group (TAG) for SDG 4.1.1.a, participants emphasized the importance of ensuring that different foundational learning assessments are aligned with benchmarks in a way that allows for valid comparisons—both with established Minimum Proficiency Levels (MPLs) and among the assessments themselves. The TAG noted the need to better define the difficulty level of newer assessments like the Early Grade Reading Assessment (EGRA), Foundational Learning Module (FLM), and People’s Action for Learning (PAL) Network assessments. While the Assessment for Minimum Proficiency Levels-a (AMPL-a) has achieved comparability through pairwise linking, newer assessments rely on benchmarks such as the percentage of correctly answered comprehension or numeracy questions. However, without specifying the difficulty of test items, these benchmarks may lack meaningful comparability.

Calibrating difficulty levels is essential to ensure that countries receive realistic performance results—avoiding undue discouragement from poor scores from difficult tests or false confidence from overly easy tests. Furthermore, results from newer assessments should be broadly comparable to those from established assessments like the Program for the Analysis of Educational Systems of CONFEMEN (PASEC) and Regional Comparative and Explanatory Study (ERCE), which have been widely used in low- and low-middle income countries. The goal is to ensure consistency in determining the percentage of students reaching the MPL. Despite these concerns, the TAG cautioned against over-specification, as the objective is not to create a new global assessment but rather to improve the alignment and interpretability of existing ones.

1.2 Estimating Difficulty of Foundational Mathematics Tests

To better define the difficulty level of items assessing foundational constructs, it is important to consider the assessed content and item-level features. In mathematics, item-level features include both construct-relevant and construct-irrelevant elements. Construct-relevant elements include aspects of the assessed content that may vary such as the number ranges, geometric shapes, and complexity of data displays. For example, two items may assess the same early numeracy skill of adding within 30 but have different difficulty estimates based on additional construct-relevant skills. Take the items “ $23+2=$ __” and “ $19+4=$ __”. These items both assess the same skill but “ $23+2=$ __” does not require regrouping in the ones place and may be less difficult than “ $19+4=$ __”, which requires

¹ Authored by Leanne Ketterlin Geller under the guidance of Luis Crouch.



students to regroup the ones. These construct-relevant elements are important to consider when designing mathematics tests that are balanced in difficulty.

Construct-irrelevant elements include item-level and test-level features that are not associated with students' knowledge, skills, and abilities in the tested construct. For example, reading a mathematics word problem may assess students' reading skills, which are not relevant to their mathematics skills. If students' reading skills hinder their ability to understand the word problem, students' responses may not be an accurate representation of their mathematics skills. Another example may involve the quality of the printed test material; items that are printed in small font or with limited space for students to write their answers may require additional skills such as visual acuity or fine motor skills. These additional skills may impact students' ability to demonstrate their knowledge, skills, and abilities in mathematics. As such, items should be designed with minimal construct-irrelevant elements to improve the accuracy of measurement.

To provide guidance about how item-level features impact item difficulty estimates, we examined the item statistics from approximately 1,100 items designed to measure foundational mathematics constructs. Using these analyses, 50 exemplar items were written to illustrate the item-level features that impact difficulty estimates. This report describes the process of examining the items and the rationale for the exemplar items.

2. Mathematics Assessments

2.1 Assessments of Foundational Mathematics Constructs

Tests included in these analyses assessed foundational mathematics concepts as identified by the Global Proficiency Framework (GPF; see Addendum for Grade 2). To be consistent with the terms used in the GPF, this report follows the same naming convention. As such, the term “construct” represents the content areas within the five domains of mathematics. For example, within the domain of “Numbers and Operations,” there are six constructs including “Whole numbers” and “Fractions”. Within each construct, there are multiple subconstructs. For example, within the construct of “Whole numbers,” there are four subconstructs including “Identify and count in whole numbers and identify their relative magnitude” and “represent whole numbers in equivalent ways.” For each of the subconstructs, additional specificity is provided for the knowledge, and skills students should know and be able to do at each grade, from Grade 1 through Grade 9. These statements range from knowledge and skills that “partially meet global minimum proficiency,” “meet global minimum proficiency,” and “exceed global minimum proficiency.” This report targets the knowledge and skills that “meet global minimum proficiency” in Grade 2 of the GPF.

In July 2024, the UIS Director submitted a request to various countries and USAID to receive examples of foundational mathematics tests that were well-viewed by the Ministry as well as data from test administrations. The following tests and data sets were received:

Assessment	Forms or Subtests Received	Data Received
Early Grade Mathematics Assessment (EGMA)	Number Identification Quantity Comparison Missing Number Addition Level 1 Subtraction level 1 Addition Level 2 Subtraction Level 2 Word Problems	Ghana 2013 Grade 2 Ghana 2015 Grade 2 Jordan 2017 Grade 2, 3 Tanzania 2017 Grade 2* Tanzania 2022 Grade 2** Tanzania 2023 Grade 2
Early Grade Mathematics Assessment (EGMA)	Missing Number Addition Level 2 Subtraction Level 2 Relational Reasoning Spatial Reasoning	Kyrgyzstan 2021 Grade 2 Kyrgyzstan 2024 Grade 2
UNICEF Foundational Learning Module (FLM) 2.0	104 items	Proprietary field tested from Kenya 2024
AMPL-ab	8 Test Booklets	Publicly available data from 2023
International Common Assessment of Numeracy (ICAN)	Set 1 Set 2	No data were obtained

* Number identification not administered

**Addition and Subtraction level 1 not administered

We received some tests that sampled knowledge and skills from grades 4 and beyond. These tests included the USAID administration of a modified EGMA in Kyrgyzstan 2021 and 2024 in Grade 4 as well as an EGMA administration in Tanzania 2022 and 2023 in Grade 4. AMPL-ab also included items that assessed knowledge and skills at Grade 4 and beyond.

2.2 Content Alignment of the Tests of Foundational Mathematics Constructs

Items from the mathematics tests were examined to identify the alignment with the GPF. Items were coded for the construct and subconstruct assessed. The targeted grade was determined by examining the grade at which the assessed knowledge and skills aligned with the “meets minimum global proficiency” level. The final spreadsheet included 1,098 items.

3. Analyses and Results

3.1 Item Analyses

AIR analyzed the EGMA data and reported item difficulty by calculating the p -value (the proportion of students who responded correctly to an item out of the total number of students who attempted the item). It is important to note that p -values are sample dependent and thus may vary based on the skills of the sample who responded to the items. For example, a sample of respondents with strong skills in solving operations using whole numbers (subconstruct N1.3) will likely have a larger proportion of students responding correctly than a sample with emerging skills. The item p -values from the former sample will be higher (indicating lower difficulty) than the p -values from the latter sample (indicating higher difficulty).

To illustrate the issue of sample dependency, Table 1 compares the median p -values for Grade 2 students from three countries who took similar subtraction tests. Also included are socio-economic variables that provide insights into the country contexts.


Table 1. Median p -value for subtraction items for Grade 2 students by socio-economic variables in three countries

Country	Item Statistics		Socio-economic Variables*		
	Median p -value	Median item-total correlation	GDP per capita (current US\$)	Under-Five Mortality Rate	Secondary school enrollment (% gross)
Ghana (2013, 2015)	0.31	0.61	1850	58	64
Jordan (2017)	0.79	0.56	4082	17	89
Tanzania (2017, 2023)	0.50	0.69	1075	48	28

*Source: GDP per capita (current US\$), under-five mortality, and secondary school gross enrollment ratio: downloaded from World Bank World Development Indicators; all data are averaged from two years before and two years after the application of the EGMA assessments.²

The median p -value for countries with lower socio-economic indicators (e.g., lower GDP per capita, higher under-five mortality rate) is lower than the median p -value for Jordan, whose socio-economic indicators are more robust. In other words, a smaller proportion of

² The secondary school enrollment (% gross), not the primary school enrollment ratio was chosen because by now most countries have reached nearly 100% primary gross enrollment, so there would not be enough variance in the ratio to make it a useful indicator of social development.



students in Ghana and Tanzania responded correctly to similar subtraction items than did students in Jordan. These data indicate that subtraction items are more difficult for Grade 2 students in Ghana and Tanzania than for students in Jordan. These findings may be a result of variability in the instructional environment or educational opportunities across these contexts. For all countries, the median item-total correlations are above the threshold of 0.20, indicating that the items are measuring a similar construct.

AIR also reported the item-total correlations as an indicator of item quality. Item-total correlations point to how well the item aligns with the overall construct being measured. Items with low item-total correlations may include construct-irrelevant elements or may be assessing an unrelated construct.

Data obtained from UNICEF's FLM 2.0 items were from a pilot test administered to students in grades 2-4 in Kenya in 2024. To be consistent with the analyses provided by AIR, p-values were used as the indicator of item difficulty and the item-total correlations as an indicator of item quality.

Data from AMPL-ab were obtained from a publicly available data set from a sample of students in grades 3-8. Because the sample extended beyond the targeted grade, these data were not used to evaluate item difficulty and item quality. Instead, the items were used to evaluate item-level features that may lead to accurate measurement of students' knowledge and skills in the assessed content.


No data were available for the ICAN items.

3.2 Results: Item Difficulty by Subconstruct

The purpose of these analyses was to examine the range of item difficulty estimates by subconstruct of the GPF. With this information, exemplar items were written that align with known difficulty estimates. These exemplars serve as guidelines for future test development efforts in foundational mathematics constructs.

Prior to analyzing the item difficulty statistics, items were excluded for two reasons. First, items with item-total correlations below 0.20 were excluded because this could be an indicator of misalignment with the overall latent construct of foundational mathematics knowledge and skills. Second, items that assessed subconstructs beyond grade 3 of the GPF were removed.

Table 2 describes the outcomes of these analyses by subconstruct using the labels consistent with the GPF (e.g., A1.1 is associated with subconstruct "recognize, describe, extend, and generate patterns" in the domain of Algebra in the GPF). The total number of items assessing each subconstruct by grade are reported. A majority of the items across




these tests assess N1.1 (identify and count in whole numbers, and identify their relative magnitude) and N1.3 (solve operations using whole numbers). The fewest number of items ($n=5$) assess M2.2 (solve problems involving time).

Item difficulty statistics are reported by subconstruct. The lowest p -value represents the most difficult items (fewer grade 2-3 students responded correctly). For example, a p -value of 0.49 indicates that 49% of the sample of students who attempted the item responded correctly. The highest p -value represents the least difficult items (a larger number of grade 2-3 students responded correctly). For example, a p -value of 0.89 indicates that 89% of the sample of students who attempted the item responded correctly. Because the subconstructs were assessed by items beyond grade 3, we also calculated the average p -value for items across all grades.

The most difficult subconstruct for students in grades 2-3 is S1.1 (Retrieve and interpret data presented in displays) with a lowest p -value of 0.05 and an average p -value of 0.10. The least difficult subconstruct for students in grades 2-3 is G1.1 (recognize and describe shapes and figures) with an average p -value for students in grades 2-3 of 0.90.

Table 2. Item Difficulty Estimates by Subconstruct

Mathematics Subconstruct in Grade 2 of the GPF	Item Count: Content by Grade			Item Difficulty Statistics (p -values)			
	1	2	3	Lowest p - value for G2-3 (most difficult)	Highest p - value for G2-3 (least difficult)	Average p - value for G2-3	Average p - value across all grades
A1.1: Recognize, describe, extend, and generate patterns		6		0.49	0.89	0.61	0.60
G1.1: Recognize and describe shapes and figures	1	6		0.46	0.91	0.90	0.81
G2.1: Compose and decompose shapes and figures		6		0.25	0.80	0.57	0.55
G3.1: Describe the position and direction of objects in space		7	1	0.41	0.75	0.65	0.63
M1.1: Use non-standard and standard units to measure, compare, and order	1	5	1	0.26	0.70	0.44	0.46
M2.1: Tell time		7		0.61	0.74	0.68	--
M2.2: Solve problems involving time		5		0.15	0.71	0.41	0.44
M3.1: Use different currency units to create amounts		6		0.36	0.83	0.59	--
N1.1: Identify and count in whole numbers, and identify their relative magnitude	56	104	75	0.04	0.99	0.56	0.63



N1.2: Represent whole numbers in equivalent ways	2	4	1	0.44	0.89	0.70	0.63
N1.3: Solve operations using whole numbers	99	150	41	0.02	0.98	0.45	0.45
N1.4: Solve real-world problems involving whole numbers	6	13	1	0.13	0.93	0.43	0.44
S1.1: Retrieve and interpret data presented in displays	1	8	1	0.05	0.60	0.10	0.35



4. Exemplar Items by Subconstruct

To design exemplar items for each subconstruct by difficulty range, a test blueprint was written. Table 3 identifies the range of difficulty estimates (in p-values) for each subconstruct. Within each cell is an item identification code that corresponds to the Item ID in the item spreadsheet. The Item ID references the subconstruct (e.g., A1.1, G1.1), the approximate difficulty (low-, middle-, or high-range p-values), and the sequential number within the difficulty range (e.g., 1-3). Cells without Item IDs indicate that no generalizable pattern of item difficulty within that range was observed. As such, exemplar items could not be written.

The test blueprint signals the most common clusters of item difficulty estimates from the existing foundational mathematics tests. For example, for A1.1, most items had difficulty estimates that fell into three clusters with p-values in the ranges of $p = 0.40-.49$, $p = 0.50-.59$, and $p = 0.80-.89$. Using the item content for these existing items, a new item was then written to illustrate the item-level features for this difficulty range. For example, three items were written for A1.1 (recognize, describe, extend, and generate patterns) that correspond with the three ranges of item difficulty estimates. The difference in difficulty of the items could be attributed to the complexity of the pattern and the missing pattern unit. The most difficult item, A1.1_Mid1, is designed with a difficulty of approximately 0.40-.49 and includes a pattern with three shapes in the pattern unit with 2 repeating shapes. Students need to identify the missing pattern unit from the middle of the pattern. The least difficulty item, A1.1_Hi2, is designed with a difficulty of approximately 0.80-.89 and includes a pattern with three shapes in the pattern unit with no repeating shapes. Students need to identify one missing shape from the end of the pattern.

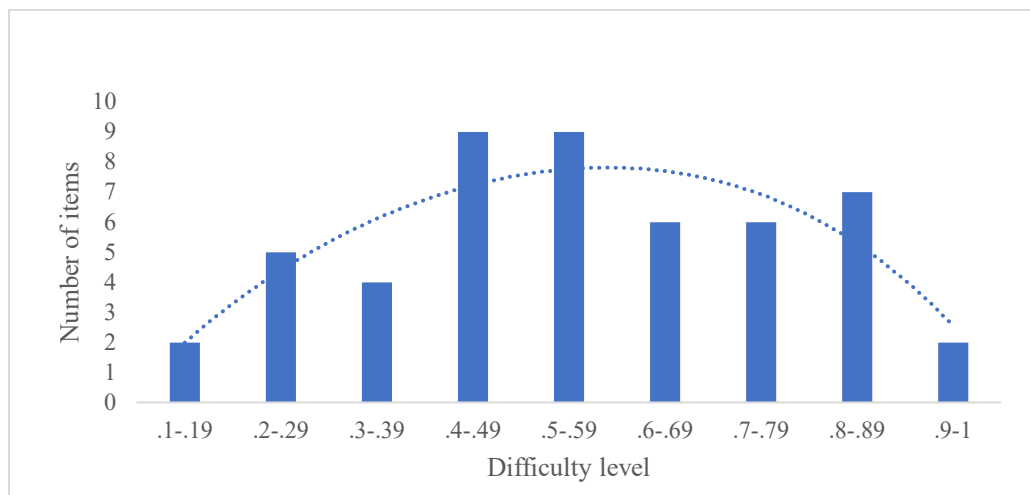
Table 3. Test Blueprint

	Low p-values (most difficult)			Mid-range p-values (medium difficult)			High p-values (least difficult)			N
	0.10-0.19	0.20-0.29	0.30-0.39	0.40-0.49	0.50-0.59	0.60-0.69	0.70-0.79	0.80-0.89	0.90-1	
A1.1: Recognize, describe, extend, and generate patterns				A1.1_Mid1	A1.1_Mid2			A1.1_Hi1		3
G1.1: Recognize and describe shapes and figures				G1.1_Mid1				G1.1_Hi1	G1.1_Hi2	3
G2.1: Compose and decompose shapes and figures		G2.1_Lo1		G2.1_Mid1		G2.1_Mid2		G2.1_Hi1		4
G3.1: Describe the position and direction of objects in space				G3.1_Mid1		G3.1_Mid2	G3.1_Hi1			3
M1.1: Use non-standard and standard units to measure, compare, order		M1.1_Lo1			M1.1_Mid1		M1.1_Hi1			3
M2.1: Tell time						M2.1_Mid1	M2.1_Hi1			2
M2.2: Solve problems involving time		M2.2_Lo			M2.2_Mid1					2

M3.1: Use different currency units to create amounts			M3.1_Lo1		M3.1_Mid1			M3.1_Hi1		3
N1.1: Identify and count in whole numbers, and identify their relative magnitude	N1.1_Lo1	N1.1_Lo2	N1.1_Lo3	N1.1_Mid1	N1.1_Mid2	N1.1_Mid3	N1.1_Hi1	N1.1_Hi2	N1.1_Hi3	9
N1.2: Represent whole numbers in equivalent ways				N1.2_Mid1	N1.2_Mid2		N1.2_Hi1	N1.2_Hi2		4
N1.3: Solve operations using whole numbers	N1.3_Lo1	N1.3_Lo2	N1.3_Lo3	N1.3_Mid1	N1.3_Mid2	N1.3_Mid3	N1.3_Hi1	N1.3_Hi2		8
N1.4: Solve real-world problems involving whole numbers			N1.4_Lo1	N1.4_Mid1	N1.4_Mid2					3
S1.1: Retrieve and interpret data presented in displays				S1.1_Mid1	S1.1_Mid2	S1.1_Mid3				3
Total	2	5	4	9	9	6	6	7	2	50

The distribution of item difficulty estimates is approximately balanced with a concentration of items around the mid-point of the range. This satisfies the recommendation of the TAG. Figure 1 provides a graphical description of the distribution of the items across the difficulty levels.

Figure 1: Distribution of items by difficulty level




5. Conclusions

5.1 Implications for Future Test Development

A detailed analysis of item difficulty estimates for existing items assessing foundational mathematics constructs at Grade 2 revealed distinct patterns in item-level features that influence variability in difficulty. These findings were leveraged to design exemplar items, which, in turn, can serve as valuable tools for guiding the development of future assessments or refining existing ones aligned with the GPF in Grade 2.

For test developers working on new foundational mathematics tests, both the sample test blueprint and the exemplar items offer essential guidance in designing a test that is well-balanced in both content and difficulty. The test blueprint provides a structured framework that outlines the observed distribution of items across subconstructs and difficulty levels, ensuring comprehensive coverage of the GPF. Test developers can strategically target specific subconstructs and adjust item difficulty to achieve the desired balance. By referencing the exemplar items, test developers can gain insights into item-level features that correspond to different difficulty levels, allowing for more precise item design.

Using these exemplar items as models can also streamline and expedite the test development process. Since the item difficulty of these items are already established, developers can reduce the need for extensive development of untested items, saving time



and resources. Once new items are drafted, they still need to undergo rigorous evaluation for content alignment and psychometric properties before being included in operational tests.

Additionally, these analyses and exemplar items can guide revisions to existing foundational mathematics tests. If test developers determine that their existing test does not adequately cover the full range of difficulty levels necessary for a balanced assessment, they can use the exemplar items as templates for creating new items that enhance test coverage. As with newly developed tests, any revisions to existing assessments require careful psychometric evaluation and content review before implementation.

5.3 Implications Associated with Sample Dependence

It is important to reiterate that the difficulty estimates used in this report – specifically, p -values – are sample dependent. As a result, the exemplar items may perform differently when administered to actual students than is predicted by their difficulty estimates. There are two key implications that should be explored.

First, as previously described, p -values are normatively derived, meaning they reflect the performance of a specific sample. This can be advantageous when developing tests intended for particular contexts, where results are not meant to be generalized beyond that setting. When designing a test for a specific sample, it is important that the items align with the knowledge, skills, and abilities of that population. If the items are too easy, a ceiling effect may occur, limiting insights into the variability of examinees' abilities. Conversely, if the items are too difficult, a floor effect may result, similarly restricting the ability to accurately measure the targeted knowledge and skills effectively. Because p -values are influenced by the sample's abilities, they provide useful information on whether the items – and the overall test – are appropriately challenging. For test results to be meaningful and accurately interpreted, the difficulty of both individual items and the overall test should be well matched to the specific sample.

Second, in contrast to situations describe above in which decisions are *not* intended to generalize beyond a specific context, it is not tenable to use test scores from items calibrated with locally derived p -values to compare outcomes across contexts or across time. Locally derived p -values may lead to significant variations in test difficulty when assessments are developed separately in different educational contexts. For example, if two regions with distinct instructional practices and learning opportunities design their own assessments using locally calibrated p -values, the resulting tests may not be comparable in terms of difficulty. Consequently, when test scores are intended for cross-context comparisons or comparisons over time, items should be calibrated using Item Response Theory (IRT) modeling or with samples that represent a broad spectrum of the targeted population.




5.4 Limitations and Considerations

The foundational mathematics tests used to generate the item difficulty estimates and exemplar items were primarily focused on the Numbers and Operations domain. Few items were available that represented the other four domains in mathematics. As a result, there is limited evidence on how item-level features influence difficulty estimates for many subconstructs outside of Numbers and Operations. This gap in data is significant because it limits the generalizations we can make about items across domains. To continue providing meaningful guidance for the development and refinement of foundational mathematics tests, more data from a wider range of tests are needed. Expanding the dataset to include items from all domains will allow a deeper analysis of item difficulty patterns, thereby improving the scope of the sample test blueprint and the range of exemplar items.

Although data were obtained from various countries, the sample of students may not be representative of the broader population.

Item-response type may impact difficulty estimates. The tests used for these analyses included selected-response and constructed-response type items. As additional data are available from other foundational mathematics tests, these analyses could be repeated by item-response type.



TECHNICAL DOCUMENTATION SUPPORTING
THE EXEMPLAR FOUNDATIONAL MATHEMATICS ITEMS

Email:

uis.information@unesco.org

uis.director@unesco.org

uis.unesco.org

[@UNESCOstat](https://twitter.com/UNESCOstat)