AI-Based Classification of TIMSS Items for Framework Alignment

Ummugul Bezirhan and **Matthias von Davier**
TIMSS & PIRLS International Study Center at Boston College
{bezirhan, vondavim}@bc.edu

Abstract

Large-scale assessments rely on expert panels to verify that test items align with prescribed frameworks, a labor-intensive process. This study evaluates the use of GPT-4o to classify TIMSS items to content domain, cognitive domain, and difficulty categories. Findings highlight the potential of language models to support scalable, framework-aligned item verification.

1 Introduction

International large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS) play a critical role in monitoring educational outcomes across diverse systems. The validity argument of such assessments lies in the rigorous alignment of test items with the underlying assessment framework, which defines key content and cognitive domains that the assessment purports to measure. TIMSS assessment development is guided by the principles of Evidence-Centered Design (Mislevy et al., 2003), ensuring that each item serves as meaningful evidence for the targeted constructs. This process involves multiple rounds of expert review and collaboration with participating countries to verify item alignment and maintain the validity of measurement across contexts.

While effective, this expert-driven validation process is labor-intensive and time-consuming, particularly in the context of ongoing item development and reuse. As AI technologies continue to evolve, they offer new ways for automating or supporting some of these processes. One such approach is the use of large language models (LLMs) for automated item classification. If reliable, these tools could significantly reduce the burden on subject matter experts, streamline assessment development cycles, and enhance scalability without compromising psychometric quality.

This study explores the potential of GPT-4o to perform classification of TIMSS 2019 mathematics items. Specifically, we evaluate the model's ability to assign items to their appropriate content domain, cognitive domain, and difficulty level, based on the given TIMSS assessment framework. The items have already been reviewed and validated by expert panels and are used operationally, their classifications can be considered reliable benchmarks.

To assess alignment, AI-generated classifications are compared against expert-coded categories, analyzing agreement patterns and identifying systematic divergences. For difficulty, we define three difficulty regions using percent correct values derived from empirical item performance data and evaluate the model's capacity to approximate these classifications. The findings of this study contribute to ongoing discussions about the role of AI in assessment development and offer preliminary evidence on the feasibility of LLMs as tools to support item verification within established assessment frameworks.

2 Background

Construct validity has long been a central concern in educational assessment, particularly in international large-scale assessments such as TIMSS. A key aspect of evidence for validity is the alignment between test items and the assessment framework, that is the extent to which each item's

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 134–141
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

content and cognitive demands reflect the intended constructs of the study. Alignment in the context of ILSAs supports meaningful score interpretation, facilitates cross-national comparability, and provides assurance that assessment inferences are based on systematically defined learning goals. This also helps minimize the construct irrelevant variances.

Foundational work on test design and validity, such as Messick's (1990) unified validity framework and ECD of Mislevy et al. (2003), emphasizes that the validity argument must include an explicit evidentiary chain connecting item features to well-articulated domain models. Alignment research is one way to establish this chain by evaluating the connection between testing, content standards, and instruction. If these components work together to deliver a consistent message about what should be taught and assessed, students will have the opportunity to learn and to truly demonstrate what they have achieved (Martone & Sireci, 2009). Systematic alignment studies therefore provide critical priori evidence that the assessment operationalizes its framework as intended, thereby supporting the overall construct-validity argument.

In the context of TIMSS, alignment involves a multistep process where items are reviewed, refined, and approved by subject matter experts, ensuring they adhere to content domains, cognitive processes and intended difficulty levels. While this process is foundational to the psychometric integrity of the assessment, it is also resource-intensive and difficult to scale given growing item pools and evolving frameworks.

To address these challenges, researchers have explored the use of computational methods to support or automate parts of the alignment process. Advances in natural language processing (NLP) have opened new possibilities for supporting alignment through semantic analysis of item texts. Recent studies (e.g., Butterfuss & Doran, 2024; Camilli, 2024; Camili & Suter, 2024) have demonstrated that embedding-based similarity metrics can successfully identify meaningful relationships between standards and item specifications. Such methods have been used in alignment studies involving the Common Core State Standards and NAEP, showing that NLP techniques can reproduce many expert classifications through clustering or regression models. While promising, these approaches often rely on static sentence embeddings and do not fully capture the contextual reasoning that human experts employ when classifying items.

Building on this prior work, the current study investigates the use of a large language model, GPT-4o, to perform classification of TIMSS mathematics items in alignment with the given TIMSS framework. By incorporating the full descriptive language of the framework into the prompt through a structured prompt engineering approach that dynamically loads framework specifications from a framework focused database, this method allows complete content domain descriptions, cognitive skill definitions, and difficulty level characteristics specific to each TIMSS assessment year and grade level. Unlike previous efforts that focus on pairwise similarity, this dynamic framework-informed prompting strategy offers a scalable, interpretable, and multidimensional approach to item classification, potentially streamlining alignment procedures while preserving the integrity of the assessment development process.

3 Methods

3.1 Data Source

This study uses a sample of mathematics items from TIMSS 2019 for Grade 4 and Grade 8 assessments. All selected items were previously reviewed and validated by expert panels convened by TIMSS and PIRLS International Study Center and successfully field tested. Each item includes a final assigned content domain, cognitive domain, and empirical difficulty estimate based on percent correct values from operational test data.

The study includes all newly developed items introduced in the TIMSS 2019 cycle. For items containing images, diagrams, or graphs, the GPT-4o model via the OpenAI API was used to generate descriptive captions, allowing for the full item set to be processed in text-based analyses. In each TIMSS cycle items are selected to ensure coverage across a range of content topics (e.g., number, algebra, life science), cognitive domains (knowing, applying, reasoning), and difficulty levels. The complete dataset initially consisted of 286 items. However, items split into multiple parts (e.g., a, b, c sub-items) were excluded from the classification analysis to avoid duplication and ensure consistency in unit of analysis. After this filtering, the final analytic sample comprised 217 items.

Table A1 shows the item distribution by each category in Appendix A.

3.2 Framework Representation and Prompt Design

To support classification by the language model, we constructed structured prompts embedding full descriptions of TIMSS framework dimensions. TIMSS 2019 Assessment Framework (Mullis & Martin, 2017) served as a primary source for content domain definitions, cognitive domain descriptions, and difficulty-level guidance.

A custom framework database was built utilizing PDF descriptions of the frameworks to dynamically retrieve definitions relevant to the grade level and subject of each item. Prompts followed a template-based structure that presented:

- The item content
- The TIMSS subject, grade level, and year
- Full framework definitions for the content domains
- Full framework definitions for the three cognitive domains
- Empirical guidance for difficulty classification

Condition	Description	Examples	CoT
Zero-shot (ZS)	Framework definitions + item only	None	No
Zero-shot CoT (ZS-CoT)	Adds “Think step by step” instruction	None	Yes
Few-shot (FS)	Adds one example per cognitive × difficulty cell	9-10	No
Few-shot CoT (FS-CoT)	Adds one example per cognitive × difficulty cell and CoT reasoning	9-10	Yes

Table 1: Prompting Conditions

An example of the prompt is given in Figure A1 in Appendix A.

In addition to aligning with the official TIMSS framework, this study examined how prompt design strategies influence the language model’s classification performance across three target dimensions: content domain, cognitive domain, and difficulty level.

Recent advances in natural language prompting have shown that model performance can be improved by structuring reasoning and task representation within the prompt itself. Two key strategies examined in this study are Chain-of-Thought (CoT) prompting and meta-prompting. CoT prompting encourages the model to generate step-by-step reasoning before producing a final answer, supporting tasks that involve multi-step inference or abstract judgment (Wei et al., 2022). This approach is particularly relevant for educational item classification tasks, where judgments such as cognitive demand and difficulty are often nuanced and require the model to simulate student and/or expert thinking.

Building on this, meta-prompting involves instructing the model on how to perform the task itself by embedding structured guidelines directly into the prompt (Reynolds & McDonell, 2021; OpenAI, 2024). In more advanced forms, meta-prompts may enable models to critique or revise their own instructions or those provided by users (Ye et al., 2023). Recent work has further enhanced this approach by labeling individual reasoning steps and implementing step-aware verifiers, which assess each step’s contribution to the final decision (Li et al., 2023).

To evaluate the influence of prompt structure on classification performance, the study implemented four prompt conditions shown in Table 1.

3.3 Model and Classification Procedure

We used GPT-4o, accessed via OpenAI’s API, as the large language model for classification. Each item prompt was submitted independently, and the model’s textual response was parsed to extract predicted content domain, cognitive domain, and difficulty level. A post-processing script was applied to standardize terminology and correct minor inconsistencies such as the content domain in grade 4 is ‘Measurement and Geometry’ but the model specified the items as ‘Geometry’ or ‘Measurement’, those were counted as ‘Measurement and Geometry’.

The classification process was fully unsupervised; no labeled training data or fine-tuning was used. All responses were generated using temperature = 0 to maximize determinism and reproducibility.

Model performance was evaluated by comparing model’s predicted content and cognitive domain classifications to expert-assigned labels. Content and cognitive domain accuracies reflect the proportion of exact matches between the model's predictions and the official domain labels. Difficulty classification was evaluated against empirical difficulty levels derived from operational data. Specifically, items were categorized as Easy, Medium, or Hard based on their percent-correct values, using Easy (>60%), Medium (30-60%), and Hard (<30%). The model's predicted difficulty level was considered correct if it matched the empirically derived category for each item. Cohen's kappa coefficients were also calculated to account for chance agreement. Additionally, misclassifications were analyzed qualitatively to identify systematic patterns of divergence.

4 Results

Classification Performance

Classification performance across prompting conditions is summarized in Table 2. Content domain classification demonstrated consistently high performance, with all prompting conditions exceeding 94% accuracy and kappa values above 0.92, indicating substantial agreement beyond chance. FS-CoT achieved the highest accuracy (94.9%) and kappa (0.933), reflecting the model's strong ability to differentiate TIMSS content domains. In contrast, classification accuracy for the cognitive domain showed more variation, ranging from 60.4% to 64.1% and kappa values between 0.382 and 0.438. The FS-CoT condition yielded the highest accuracy, followed by ZS baseline and FS. Kappa values across these conditions suggest fair to moderate agreement with expert labels, indicating that while the model captures meaningful cognitive distinctions, it does so with less precision than in the content domain.

Difficulty classification, while the most challenging of the three dimensions, showed improvement over previous iterations. Accuracy scores ranged from 44.2% to 49.8%, and all conditions resulted in positive kappa values, indicating better-than-chance agreement. ZS-CoT led in both accuracy and agreement, though overall performance remained modest, highlighting the inherent complexity of predicting empirically derived difficulty levels. Grade level analysis revealed consistently stronger model performance for Grade 4 items across all classification dimensions. For content domain classification, Grade 4 items achieved exceptional accuracy scores ranging from 96.9% to 97.7%. Grade 8 content domain performance, while lower, remained strong with accuracy scores from 91.0% to 92.3%. A similar pattern was also observed in cognitive domain classification. Grade 4 accuracy ranged from 62.6% to 65.0%, while grade 8 performance varied from 57.4% to 62.8%. Notably, the FS-CoT condition achieved the smallest grade-level gap in cognitive domain performance (65.0% vs. 62.8%). For difficulty classification, Grade 4 items consistently outperformed Grade 8 items across all conditions. Grade 4 difficulty accuracy ranged from 50.4% to 57.7%, with ZS-CoT achieving the highest Grade 4 performance (57.7%). Grade 8 difficulty classification proved more challenging, with accuracy scores ranging from 33.0% to 40.4%, with FS-CoT achieving the best Grade 8 performance (40.4%).

Prompt	Content Domain		Cognitive Domain		Difficulty Level	
	Acc	κ	Acc	κ	Acc	κ
ZS	94.1	0.922	62.2	0.410	44.2	0.072
ZS CoT	94.2	0.923	60.4	0.382	**49.8**	**0.134**
FS	94.1	0.923	61.3	0.397	44.7	0.074
FS CoT	**94.4**	**0.930**	64.1	0.438	48.4	0.097

Table 2: Classification Performance

Classification Patterns and Systematic Errors

Given its overall better performance across all three classification dimensions, the FS-CoT condition was selected for detailed confusion matrix analysis to understand specific classification patterns and systematic errors.

For the content domain classification, the model achieved near perfect classifications, but specific patterns emerged when analyzed by grade level (Appendix A Figures A2-A3). For Grade 4 mathematics, the model achieved perfect classification for Data and Number domains but showed some boundary confusion with Measurement and Geometry items. Specifically, 12% of Measurement and Geometry items were misclassified as both Data and Number domains, suggesting overlapping conceptual features in items involving spatial reasoning and numerical computation. For Grade 8, content domain classification revealed different boundary challenges. While Algebra, Data and Probability, and Geometry domains were classified perfectly, Number domain items showed notable confusion (75%). The primary misclassification pattern involved 21% of these items being classified as Algebra, with an additional 4% classified as Geometry. This pattern suggests that model struggles with the increasing integration of algebraic thinking into numerical context in the higher grades.

As shown in Figure A4, the FS-CoT model exhibited a strong bias toward predicting the Applying domain. While Applying items were accurately classified 84% of the time, it also attracted most misclassifications receiving 45% of Knowing and 44% of Reasoning items. Reasoning accuracy was moderate (53%) but showed substantial confusion with Applying. Very few items were confused between Knowing and Reasoning, indicating the model can generally distinguish between higher-order and basic cognitive demands but struggles to differentiate between applying procedures and engaging in mathematical reasoning.

Difficulty classification remained the most challenging task for the model, with a strong tendency toward underestimation (Figure A5 in Appendix A). Easy items were correctly classified 64% of the time and no easy items were misclassified as Hard, indicating a cautious estimation pattern. Medium items had 71% accuracy, with 26% underestimated as Easy and only 3% overestimated as Hard. This suggests the model treats difficulty as a binary decision *Easy versus Not Easy* rather than effectively distinguishing all three levels. If we collapse the difficulty to this more pragmatic *Easy vs. not Easy* decision, the accuracy jumped to 0.78. Hard items were the most frequently misclassified. This reflects a consistent failure to recognize complex mathematical or cognitive demands, particularly when such items are concise or lack surface-level cues of difficulty.

Linguistic Features of Misclassified Items

To better understand the systematic errors in difficulty classification, we examined surface features of misclassified items as shown in Table 3. We focused on textual length, numerical content, and mathematical language.

	Easy	Medium	Hard
Word count	76.8	49.9	66.1
Character count	561.4	338.9	404.5
Reasoning Verb count	0.20	0.09	0.34
Number count	19.10	12.29	10.74
Operations count	0.40	1.12	1.31

Table 3: Average Surface Features of Misclassified Items

Misclassified Easy items had the highest average word count (76.8) and character length (561.4) substantially longer than misclassified Medium (50.0 words, 338.9 characters) and Hard items (66.1 words, 404.5 characters). This suggests the model tends to get confused by textual elaboration with cognitive difficulty, overestimating the challenge of otherwise straightforward tasks. Conversely, Hard items, though shorter, were rich in mathematical content. They contained the highest density of mathematical operations (1.31 per item) and reasoning verbs (0.34 per item) yet were overwhelmingly misclassified as Medium. This indicates that while GPT-4o detects complexity, it fails to properly weight them in difficulty estimation, especially when such cues are embedded in concise text.

5 Conclusion

This study evaluated the potential of GPT-4o to perform automated classification of TIMSS mathematics items. Using a dynamic, framework-aware prompting strategy, we challenged the model to assign Grade 4 and Grade 8 mathematics items to their official content domain, cognitive domain, and difficulty categories without any fine-tuning or labeled training data.

Across all prompting conditions, model consistently provided high agreement with content-domain classifications with about 95% accuracy ($\kappa > 0.92$), and confusion matrices only showed minimal boundary issues. These results suggest that content domain classification is one area where the model can be deployed with confidence.

Model accuracy for cognitive domain classifications clustered around 62% ($\kappa \sim 0.41$). This level of agreement is consistent with prior research, including Nasstrom (2009), who reported moderate inter-rater reliability ($\kappa \sim 0.41$–0.47) among experts classifying items according to Bloom's taxonomy. Similarly, Karpen and Welch (2016) found only 46% agreement among faculty when categorizing exam questions by cognitive demand. While performance improved modestly under the FS-CoT prompting condition, error analysis revealed a systematic tendency to overclassify items into the Applying category, a middle category bias. This highlights a clear opportunity for targeted prompt engineering or probability calibration strategies.

Model performance was weakest for the three-level difficulty classification task, with accuracy around 49%. However, reframing the task as a binary classification, *Easy versus Not Easy*, yielded 78% accuracy. This is particularly notable given that prior research demonstrated limited alignment between expert predictions of item difficulty and examinee performance (e.g., Bejar, 1983; Mansoor, 2024; Wonde, 2024) with accuracy rates hovering around 50-55% even after targeted expert training (Sayin & Bulut, 2024). Moreover, Clauser et al. (2009) demonstrated that physicians involved in Angoff standard setting frequently revised their difficulty estimates to align with whichever performance statistics were presented to them, regardless of their accuracy, highlighting the inherent instability of human judgements. Taken together, these findings show that unsupervised binary screening already matches or in some cases exceeds typical human baselines.

Given this, the model could serve as a first-pass filter content tagging and binary difficulty screening could reduce the number of items requiring full panel review, freeing experts time to focus on distractor quality, fairness checks, and cross-cultural comparability. In addition, because framework definitions are pulled dynamically the same pipeline can be applied to other TIMSS cycles or entirely different frameworks (e.g., NAEP, PISA) with minimal revision.

This study has potential limitations. First, the study focused exclusively on mathematics items from the 2019 TIMSS cycle; generalizability to science items, earlier cycles, or AI-generated content remains to be investigated. Second, all analyses were conducted using text-only representations of items thus visual components such as graphs or diagrams were reduced to captions, which may have affected the model's judgments. Future studies incorporating multimodal inputs may offer a more accurate reflection of the item's full content and complexity. Third, item difficulty levels were defined based on fixed percent-correct thresholds. Future research can consider using IRT-based difficulty estimates or continuous difficulty prediction using fine-tuned LLMs.

Overall, this study shows that GPT-4o, when directed with a targeted prompting strategy, can act as a reliable co-reviewer in the early stages of test development. While current results are strongest for content classification, meaningful performance in cognitive and difficulty domains, with interpretable error patterns, suggests a promising role for AI in supporting expert workflows. Rather than aiming to replace human expertise, these tools are best positioned to augment it by reducing workload and improving the speed and consistency of assessment development.

References

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303-310.

Butterfuss, R., & Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. Educational Measurement: Issues and Practice.

Camilli, G. (2024). An NLP crosswalk between the common core state standards and NAEP item specifications. arXiv preprint arXiv:2405.17284.

Camilli, G., & Suter, L. (2024). NLP Cluster Analysis of Common Core State Standards and NAEP Item Specifications. arXiv preprint arXiv:2412.04482.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*, *46*(4), 390-407.

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. *Currents in Pharmacy Teaching and Learning*, *8*(6), 885-888.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J. G., & Chen, W. (2023, July). Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5315-5333). https://aclanthology.org/2023.acl-long.291

Mansoor, M., Imran, S., Tayyab, A., & Sarfraz, R. (2024). Expert Prediction Versus Difficulty Index Measured by Psychometric Analysis; A Mixed Method Study Interpreted through Diagnostic Judgment by Cognitive Modeling Framework. *Journal of University College of Medicine and Dentistry*, 74-80.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of educational research*, *79*(4), 1332-1361.

Messick, S. (1990). Validity of test interpretation and use.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), i-29.

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Näsström, G. (2009). Interpretation of standards with Bloom's revised taxonomy: a comparison of teachers and assessment experts. *International Journal of Research & Method in Education*, *32*(1), 39-51.

OpenAI. (2024). *Prompt generation.* https://platform.openai.com/docs/guides/prompt-generation

Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1-7). https://doi.org/10.1145/3411763.3451760

Sayın, A., & Bulut, O. (2024). The difference between estimated and perceived item difficulty: An empirical study. *International Journal of Assessment Tools in Education*, *11*(2), 368-387.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Wonde, S. G., Tadesse, T., Moges, B., & Schauber, S. K. (2024). Experts' prediction of item difficulty of multiple-choice questions in the Ethiopian Undergraduate Medicine Licensure Examination. *BMC Medical Education*, *24*(1), 1016.

Ye, Q., Axmed, M., Pryzant, R., & Khani, F. (2023). Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*. https://doi.org/10.48550/arXiv.2311.05661

A Appendix

Grade	Category Type	Category	Count
4	Cognitive Domain	Applying	72
		Knowing	53
		Reasoning	39
	Content Domain	Data	60
		Measurement and Geometry	50
		Number	54
	Difficulty	Easy	43
		Medium	87
		Hard	34
8	Cognitive Domain	Applying	54
		Knowing	41
		Reasoning	27
	Content Domain	Algebra	35
		Data and Probability	26
		Geometry	26
		Number	35
	Difficulty	Easy	14
		Medium	56
		Hard	52

Table A1: Item Distribution Across Categories

Act as an expert specializing in the TIMSS assessment framework. Your task is to simulate how students interact with a {subject_name} item, diagnose its cognitive demand, and judge its difficulty level from both an expert and a student perspective.

Analyze the given TIMSS Grade {grade} {subject_name} assessment item.

Classify this item according to the TIMSS {year} {subject_name} Framework. Use these three categories:

1. **Content Domain**: Select the main content domain from this list (use the exact name):

   {content_domains_text}

2. **Cognitive Domain**: Identify the main cognitive domain (choose exactly one: Knowing, Applying, Reasoning):

   {cognitive_domains_text}

3. **Difficulty Level**: Indicate the item's difficulty (Easy / Medium / Hard), based not only on typical student success rates but also on complexity, required reasoning, potential misconceptions, distractor strength, and student accessibility:

   {difficulty_text}

Figure A1: Prompt Structure – Zero Shot

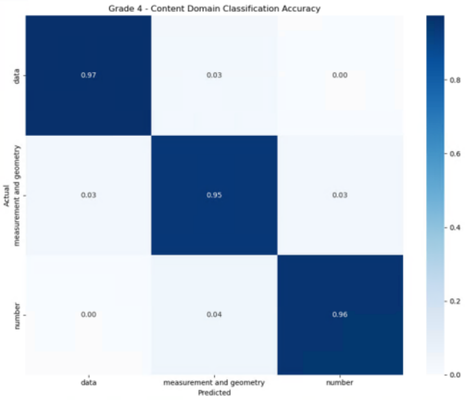

Figure A2: Grade 4 Content Domain Confusion Matrix

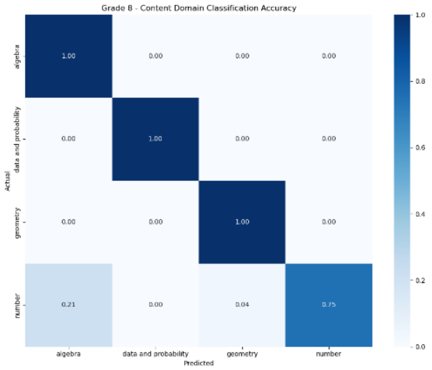

Figure A3: Grade 8 Content Domain Confusion Matrix

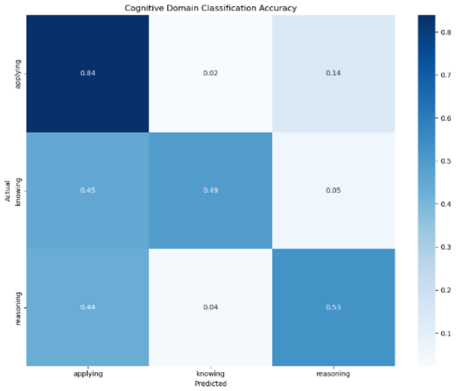

Figure A4: Cognitive Domain Confusion Matrix

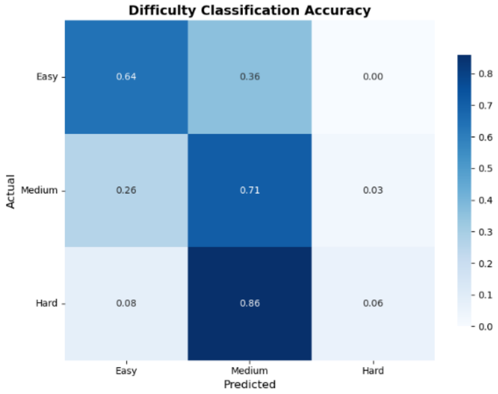

Figure A5: Difficulty Confusion Matrix