




European Journal of Educational Research

Volume 13, Issue 4, 1441 - 1453.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Writing PISA-Like Mathematics Items: The Case of Tertiary Mathematics Instructors from a State University in the Philippines

Mark Lester B. Garcia* 
Ateneo de Manila University,
PHILIPPINES

Derren N. Gaylo 
Bukidnon State University, PHILIPPINES

Catherine P. Vistro-Yu 
Ateneo de Manila University,
PHILIPPINES

Received: November 23, 2023 • Revised: January 29, 2024 • Accepted: February 19, 2024

Abstract: Mathematics test items in International Large-Scale Assessments (ILSAs) such as the Programme of International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) are nested in contexts defined in their assessment framework (e.g., the Personal, Occupational, Societal, and Scientific contexts in PISA). This study followed the item-writing activities of four tertiary mathematics instructors in the Philippines as they constructed context-based mathematics items. They were tasked to write four items each, following a set of specifications for PISA content and context categories. The data consisted of transcripts from the focus-group discussion which was conducted days after the task. The transcripts were then analyzed using thematic analysis. The results of this study showed that the phenomenon of item-writing in the context of writing PISA-like mathematics items had two themes: the phases of item-writing and the dimensions of item-writing. Findings showed that the respondents struggled to find realistic contexts and that they engaged in a problem-solving task likened to solving a puzzle as they attempted to satisfy the content, context, and process categories in the table of specifications (TOS). This study contributes to filling in the research gap on item-writing activities, particularly those of mathematics teachers in the Philippines- a country whose recent mathematical performance in the PISA 2018, TIMSS 2019, and PISA 2022 was nothing short of dismal.

Keywords: *Context-based math items, item-writing, mathematical literacy, PISA mathematics assessment framework, PISA-like mathematics items.*

To cite this article: Garcia, M. L. B., Gaylo, D. N., & Vistro-Yu, C. P. (2024). Writing PISA-like mathematics items: The case of tertiary mathematics instructors from a State University in the Philippines. *European Journal of Educational Research*, 13(4), 1441-1453. <https://doi.org/10.12973/eu-jer.13.4.1443>

Introduction

Over the past decade, the Philippines has actively participated in International Large-Scale Assessments (ILSAs) through the Department of Education's (DepEd) Bureau of Education Assessment initiative. The country's first-ever participation in the Programme of International Student Assessment (PISA) was in 2018. Moreover, it participated in the Trends in International Mathematics and Science Study (TIMSS) as early as 1995. It joined the first Southeast Asia Primary Learning Metrics (SEA-PLM) administration in 2019. These align with DepEd's goal of monitoring the implementation of the K to 12 program (DepEd, 2019), a recently implemented curricular reform reorganizing the secondary education program in the Philippines.

PISA and TIMSS are globally recognized ILSAs, as they are used in comparative studies spanning various countries and educational systems. Together with SEA-PLM, these ILSAs test students' knowledge and skills in different content domains. PISA measures the mathematics literacy, scientific literacy, and reading literacy of 15-year-old students (Organisation for Economic Cooperation and Development [OECD], 2019); TIMSS measures mathematics and science curricular content learned among Grade 4 and Grade 8 students (Mullis et al., 2016); while SEA-PLM measures the reading literacy, writing literacy and mathematics literacy of Grade 5 pupils (United Nations Children's Fund [UNICEF] & Southeast Asian Ministers of Education Organization [SEAMEO], 2019). As revealed by the reports based on PISA 2018, TIMSS 2019, and SEA-PLM 2019 results, students from the Philippines have consistently ranked bottom in terms of performance in mathematics (International Association for the Evaluation of Educational Achievement, 2020; Schleicher, 2019; UNICEF, 2021) compared to their international counterparts. Filipino learners also lag behind their regional

* **Corresponding author:**

Mark Lester B. Garcia, Ateneo de Manila University, Philippines. ✉ mark.garcia@student.ateneo.edu



counterparts in Southeast Asia, with the Philippines being only ahead of Laos in terms of average scores (United Nations Children's Fund, 2021).

Espinosa et al. (2023) discuss in great detail the performance of Filipino learners in such ILSAs. Whether the test-takers were Grade 4 students (TIMSS), Grade 5 students (SEA-PLM), Grade 8 students (TIMSS), or 15-year-old students (PISA), it was apparent that the average literacy scores for the Philippines across all domains reading, mathematics and science were significantly lower than the benchmark score (which in the case of PISA is the average of the scores of participating OECD member countries). In terms of skills, Espinosa et al. (2023) highlighted that a significant proportion of test-takers could only demonstrate basic skills indicated in the lowest proficiency levels, such as understanding simple sentences, interpreting and recognizing how simple situations can be modeled mathematically, and applying basic science knowledge to detect or identify explanations of scientific phenomena; and only around 1% of the test-takers were able to attain the highest levels of proficiency.

The country's performance trend shows no evidence of even the slightest improvement, as seen in the results of the latest cycle of PISA, which was administered in 2022. The Philippines remains among the lowest performing countries, and its average scores only changed marginally (Chi, 2023). While the Philippines' results from ILSAs remain dismal for the foreseeable future, participation in such standardized assessments remains necessary and crucial as these provide highly credible data that allow the country to benchmark with comparable educational systems with better-performing students. Kuger and Klieme (2016) explained that results from ILSAs can be used to derive recommendations for policy or practical purposes in educational effectiveness research. At the same time, data from these results enable the DepEd to periodically conduct an internal evaluation of academic policies, specifically assessment-related policies at the national and regional levels.

Several studies in the Philippines have attempted to uncover relationships between student performance in ILSAs and related variables. Examples are those of Bernardo (2021), whose work zoned in on socioeconomic and psychological factors, and of Orbeta et al. (2020), who investigated variables from different perspectives – individual, family, and school levels. Additionally, Lapinid et al. (2022) conducted a quantitative analysis of the Philippine data from the PISA 2018 results in an attempt to identify the most important predictors of mathematical performance among non-cognitive factors (e.g., physiological and socioeconomic). The focus of such studies on student-centric factors is a testament to the growing neglect of the exploration of teacher-related variables in students' performance in ILSAs.

As teachers have direct contact with students during the teaching-learning cycle in the classroom, teachers play a huge role in the success of students' learning. As highlighted in the results of the meta-analysis conducted by Kyriakides et al. (2013), teaching-related factors such as lesson structuring, questioning, and assessment had significant positive effects (with moderate effect sizes) on students' learning outcomes in the primary and secondary education level. Upon analyzing the Philippines' data from PISA 2018, Haw et al. (2021) found that teaching practices that support students' basic psychological needs positively predicted student reading achievement.

Among the teacher-related variables linked to student performance in ILSAs, however, much less attention was devoted to the teachers' assessment practices. This is rather unfortunate given that the skillful construction and use of tests can substantially improve the quality of student learning (Popham, 2011). As a matter of fact, the Commission on Higher Education (CHED) in the Philippines sought to include curriculum development and learner assessment specialists in the proposed expansion of the Technical Panel for Teacher Education as part of their measure in addressing the country's problem of low international assessment ranking (Montemayor, 2023).

Teacher assessment literacy, or the set of assessment-related knowledge and skills that teachers need to measure student achievement (Xu & Brown, 2016), has been known to have an influence on students' learning outcomes (Al-Bahlani, 2019; Elshawa et al., 2016; Hailaya, 2014; Mellati & Khademi, 2018). This is because assessment literacy helps teachers gather more accurate information about students' learning from assessments (Lian et al., 2014). Given that constructing test items is an inherent part of teachers' assessment knowledge and skills; and that students' exposure to assessments is mainly confined to their test-taking experience within the classroom, the researchers thus deem it worthy to pivot some of the attention to how teachers construct test items.

Considering this, the study explored the item-writing activities of tertiary mathematics instructors as part of a more extensive study which is the dissertation project of the primary author. One of the aims of the bigger study is to describe and identify trends among assessment practices of mathematics teachers as they construct context-based items. Moreover, it seeks to identify their potential needs when writing context-based mathematics items that must be addressed. It helps formulate possible interventions that might be useful for teachers. Specifically, this paper aimed to answer the following research questions:

1. How do mathematics teachers construct context-based mathematics items that are inspired by items from ILSAs?
2. What challenges do mathematics teachers face as they perform this assessment task?

Literature Review

Aside from instruction, assessment is central to a teacher's primary teaching duties due to the fact that performing classroom assessment-related tasks takes up a considerable chunk of their time (Stiggins, 1991, as cited in Frey et al., 2005). Additionally, results from assessment-based processes enable them to make decisions during their teaching duties at school, hence making them the "ultimate purveyors of applied measurement" (Airasian & Jones, 1993, as cited in Rodriguez & Haladyna, 2013). As professionals capable of making decisions about their learners, teachers exercise discernment in several aspects of assessment. Kalahaji and Abdullah (2016, as cited in Hanafi et al., 2020), found that a teacher's level of assessment literacy determines his or her professional judgment and interpretation.

Assessment literacy empowers teachers to pinpoint their instructional needs by being able to evaluate and use student achievement data (Khalid et al., 2021). Unfortunately, in the case of basic education teachers in the Philippines, their assessment literacy has been reported to be below the baseline level, which consists of acknowledging and responding to student progress, adjusting teaching strategies, and providing feedback to address gaps between the intended lesson and student responses (Cagasan et al., 2016). In the same vein, a similar result was echoed in the work of Napanoy and Peckley (2020) where they found that Filipino public school elementary teachers had poor assessment literacy regardless of their school type and teaching experience. The respondents' narratives revealed that they engaged in dubious assessment practices such as downloading tests from the internet and modifying these for use in classroom assessment, teaching to the test, and giving additional points in assessments for non-achievement factors such as attendance and behavior.

Piosang (2017) quantitatively measured the assessment literacy of Filipino teachers teaching English at both the secondary and tertiary levels. He found that an astounding majority of them had a classroom assessment literacy that was classified under poor level based on standards such as developing assessment methods and using assessment to determine levels of learning outcomes, among others. Meanwhile, Hailaya (2014) measured the assessment literacy levels of Filipino teachers at the elementary and secondary levels and found as well that their assessment literacy levels were low, where the teachers lacked knowledge in developing assessment methods despite knowing how to select the appropriate assessment methods.

A rather different result was observed in the study of Clores and Reganit (2020), who found that Filipino teachers in the junior high school level had mid-level assessment literacy, where they achieved the highest scores in developing assessment methods. The authors also note that math and science teachers scored higher in an assessment literacy inventory compared to their counterparts who teach social sciences, arts, and humanities. Most of the aforementioned studies, however, evaluate assessment literacy and practices of teachers from a quantitative perspective which does not offer depth in terms of the actual practices that they implement in classroom assessments.

From this brief survey of literature, there is evidence that assessment literacy is largely inadequate among Filipino teachers across various stratification levels, whether they are teaching elementary, secondary or tertiary levels, regardless of school type and teaching experience. This means that teachers have difficulty developing assessments that are appropriate for accurately measuring students' learning progress. In Philippine classrooms, traditional assessments are ubiquitous and remain the undisputed type of assessment where multiple-choice items are undoubtedly a popular choice in assessments up until the present times. Griffin et al. (2016) found that assessments in Philippine classrooms are mostly summative, and traditional assessments (e.g., true or false, short-answer, supply-type items) abound. This is not surprising given the practicality brought about by such types of items- they are easy to score, and scoring is done objectively (Geisinger & Usher-Tate, 2016).

To further examine item-writing, it must be seen as an individual activity which is highly cognitive in nature. Some studies have attempted to formulate a cognitive model of item-writing activities, and are mostly based on cognitive models of writing or written text production. Such is the work of Fulkerson et al. (2009, as cited in Fulkerson et al., 2011), where they propose three phases of item-writing when viewed from a problem-solving perspective. These phases consist of the initial representation phase, the exploration phase, and the solution phase. This model was developed further in the authors' succeeding works. The same authors observed three item-writers the following year as they performed assessment writing tasks. The participants were provided instructions for completing the task. They received templates for the task (storyboard scene template and multiple-choice item template) and a copy of the document containing science test specifications from the Minnesota Department of Education. Statements from the think-aloud recordings of the participants during the writing task were classified according to categories such as Meta-clarification, Problem definition, Missing information, Backtracking, Evaluation, Impasse, Solution satisfaction, Constraining, Relaxation, and Decomposition. The authors found that experienced writers demonstrated more forward motions in item writing compared to inexperienced writers who had more statements related to meta-clarification, missing information, or moments of impasse.

The cognitive model was then expanded to become a more comprehensive model in Fulkerson et al.'s (2011) work which now includes knowledge structures in addition to the cognitive processes in item-writing. In terms of item writing related to PISA-like math items, there are a considerable number of existing studies about them, given the widespread use of standardized testing and the influence of ILSAs. Such studies center on context-based items, like cultural factors

recommended by Memisevic and Biscevic (2022), and PISA-like mathematics items. Examples of such studies were those done by Kohar et al. (2014) and by Zulkardi and Kohar (2018), where developing prototypes of PISA-like mathematics items were documented. In both studies, task designers created items following the PISA mathematics assessment framework, where the items were then revised based on expert reviews before administering the items to students for field testing. In the former, the authors studied the PISA 2012 mathematics assessment framework prior to designing an initial item prototype. The developed items (e.g., items about temples with concentric layers of stupas) were forwarded to experts for qualitative review. Such items were then revised based on the experts' comments before administering to the students. While Kohar et al. (2014) found that the prototype items activated the fundamental mathematical capabilities of the students such as representing real-life scenarios using mathematical models and reasoning by linking information from the item with personal experience, Zulkardi and Kohar (2018) found that the task designers encountered challenges in ensuring the following desirable item qualities: authenticity of the context, accessibility of the language used, and the demand for higher order thinking skills.

An investigation of the item-writing activities of mathematics teachers as they construct traditional assessment items that are prototypes of mathematics items in ILSAs are hence warranted. Among TIMSS, PISA, and SEA-PLM, more focus is given to PISA because of the considerably higher percentage of context-based mathematics items (Ruddock et al., 2006, as cited in Close & Shiel, 2014; Wu, 2009). These kinds of items require the item writers to present a scenario for each item or a background in which the test-takers will be situating the mathematical skills that they will be applying. This is in contrast to mathematical items with no context and can be answered by a direct application of a mathematical skill (e.g., factoring, solving a linear equation, or graphing a polynomial function without any related real-life context). The item-writing literature presented involved participants who are item writers working for assessment companies and whose field of expertise is in the science subject area. Though these studies aimed to understand patterns of cognition among the participants as they engaged in writing items, writing is their primary job description, and the items they produce will be used for large-scale commercial assessments. Also, the studies on PISA-like or context-based mathematics items that were mentioned had Indonesia as their research locale, and do not expose how each task designer writes items from conceptualization until completion. On top of these, the aforementioned studies about assessment literacy and assessment practices are mostly quantitative in terms of research design. This present study thus aims to contribute to the body of knowledge on item-writing, viewing it primarily as an individual and cognitive activity, given that only a little attention is provided for the item-writing activities of teachers. Concurrently, this study also augments existing studies on PISA-like or context-based mathematics items, as such studies in the Philippines are scarce.

Methodology

Research Design

This research is a qualitative study that utilized a narrative case study approach. According to Sunday et al. (2020), this approach incorporates both case study and narrative inquiry. In this study, writing PISA-like mathematics items served as the context, while narrative became the vessel for understanding the specified case. Case study and narrative inquiry are appropriate designs for this research because together, they serve as a lens through which specific item-writing activities are magnified and deeply understood, owing to the personal experiences of the research participants (Caine et al., 2022; Duff, 2012).

Sample and Data Collection

The respondents in this study, selected through purposive sampling, consisted of four tertiary mathematics instructors from a State University in Northern Mindanao, Philippines. They were selected and recruited based on the inclusion and exclusion criteria set by the researchers. The instructors who voluntarily participated in the study, identified hereon under the pseudonyms Ben (26/M), Grace (30/F), Michaela (33/F), and Saanvi (32/F), were invited because of their academic background and teaching qualifications which are related to the field of mathematics education. They all completed a master's degree in education (primarily majoring in mathematics) and have been teaching for at least one year in their current institution. Additionally, they needed to gain experience writing PISA-like mathematics items or context-based mathematics items based on ILSA assessment frameworks. Those who declined to participate or withdrew during the timeframe of the study were excluded from the data analysis.

The research protocol, with identification number SOSEREC_22_001, underwent institutional research ethics board evaluation before the conduct of the data collection. After the ethics review, a clearance to start the implementation of the study was granted by the committee based on the application form and relevant materials submitted valid for one year. All the policies were complied with during the conduct of the research.

These four instructors underwent a 1-day seminar-workshop on writing PISA-like mathematics items which the primary author conducted via an online conferencing platform. During the first half of this training, the respondents were given an orientation seminar regarding PISA and the PISA Mathematics Assessment Framework. Then in the latter half, they were tasked to write the items using a word processing software application and all their on-screen activities were shared via screencast in the same online conferencing platform. The participants were required to write PISA-like

mathematics items that follow a given table of specifications (TOS) for content, context, and process categories. The TOS used was devised in such a way that each participant will be assigned to write four items- each corresponding to a unique content category (Quantity, Change and Relationships, Space and Shape, Uncertainty and Data) and a unique context category (Personal, Occupational, Societal, Scientific). Additionally, they were assigned to a predetermined combination of these content-context categories which corresponds to a color, as shown in Figure 1 below.

Content Category \ Context Category	Personal (25%)	Occupational (25%)	Societal (25%)	Scientific (25%)	Total
Quantity (25%)	Q1	Q1	Q1	Q1	4
Change & Relationships (25%)	Q2	Q2	Q2	Q2	4
Space & Shape (25%)	Q3	Q3	Q3	Q3	4
Uncertainty & Data (25%)	Q4	Q4	Q4	Q4	4
Total	4	4	4	4	16

*Assignments: Blue – Grace, Green – Saanvi, Yellow – Michaela, Orange – Ben

Figure 1. Item Assignments Showing the Content-Context Category Combination

For instance, since Ben was assigned orange, then he had to write four items (Q1 to Q4) with the following specifications: Q1 – Quantity and Scientific, Q2 – Change & Relationships and Personal, Q3 – Space & Shape and Occupational, and Q4 – Uncertainty & Data and Societal. For the process categories, the participants were free to assign Formulate, Employ, Interpret and Reasoning to items Q1, Q2, Q3 and Q4 in the manner they desired as long as no two items have the same process category.

A time limit of four hours was imposed so that the item writers began writing together and had to submit their items by the end of the time limit. This restriction was necessary because the more extensive study had a causal-comparative design, which compared item-writing activities of experienced and not-so-experienced mathematics teachers. Once they had turned in their items, these were forwarded to experts with experience in the administration of ILSAs in the Philippines for feedback on improving the items to make them more PISA-like in character. Once the participants were able to submit their revised items, a focus-group discussion (FGD) was conducted, where the participants were asked generic and specific questions related to their experience during and after the item-writing task.

Data Analysis

The transcripts from the FGD were then analyzed using thematic analysis. Thematic analysis is a process for identifying, examining, classifying, summarizing, and reporting themes in a set of qualitative data (Braun & Clarke, 2006). Accordingly, this method followed the following steps: familiarizing the data, generating initial codes, searching for themes, reviewing, and refining themes, and defining and naming the themes (Braun & Clarke, 2012). Applying these steps, the researchers transcribed and segmented the recordings. Participants' narratives were then coded and translated into English. The emerging themes were checked and verified by three intercoders who were experts in qualitative methodologies for reliability. After the experts' review, the themes and narratives were presented back to the participants for feedback and confirmation. This move was patterned after the work of Gaylo et al. (2020) for transparency and validity of the results.

Thematic analysis was suitable as a data analysis technique due to the fact that it is a flexible approach that provides a rich and detailed depiction of the gathered qualitative data (Braun & Clarke, 2006). In addition, it is an appropriate technique for analyzing various research participants' points of view, emphasizing parallels and discrepancies, and producing emergent themes (King, 2004).

Findings and Results

Based on the FGD transcript, the researchers identified themes that emerged from clusters of codes as discussed in this section. It consists of two themes: Phases of Item-Writing and Dimensions of Item-Writing. Each theme further consists of three subthemes.

Theme 1: Phases of Item-Writing

The first theme talks about the distinct phases in which the item-writing experiences of the participants can be classified. During the planning phase, the participants initiated their item-writing tasks, marked by their engagement in reflective behaviors at the start of the workshop. It was followed by the enacting phase, which involved explorations of potentially helpful information for the item they were currently creating and for which they were building a notable structure. Finally, the reviewing phase concludes their item-writing activity, including final revisions in preparation for submitting

their items. This phase also includes the item revisions made that were based on the feedback of Philippine experts with experience in PISA.

Subtheme 1A: Planning as a Rudimentary Precursor to Item-Writing

From the participants' accounts, it has become apparent that they needed a clear plan for approaching the item-writing task. Grace mentioned that she had set a goal of finishing the task and completing all four items- nothing more and nothing less. Meanwhile, Saanvi said that she based her items on sample items during the orientation as she did not want to exert much effort in thinking. From the video observations, it can be noticed however, that the participants constantly made consultations with materials such as the PISA mathematics assessment framework, the slides used during the seminar, a copy of the released PISA mathematics items, and the TOS template containing the assigned items as well as their content-context-process profile. These consultations occurred before and during their item-writing activities. In short, the participants applied brute force in the planning phase of their item-writing. Grace expressed that she used her instincts in writing items, and she did this by using the first thing that comes to mind when she hears the names of the content and context categories.

The planning phase serves as the foundation for item development. During this phase, it is crucial to establish clear objectives and a well-defined purpose for the assessment. At the beginning of the orientation seminar, participants were made aware that their goal was to create four items each so that all 16 items follow a predetermined TOS. Additionally, they were provided copies of pertinent documents such as item submission templates and the PISA mathematics assessment framework, which served as their guide in the item-writing task. Observing a teacher plan his or her items is difficult as this behavior is covert. Johnson et al. (2017) described their first phase of item-writing as thinking about writing, which only makes itself evident through certain observable behaviors such as typing (as shown by the respondents' screencasts). However, none of the typing behaviors shown were related to planning, as most of the respondents' planning was performed mentally.

Another covert behavior observed was that the participants struggled to align their items with the content, context, and process categories stipulated in the PISA mathematics assessment framework (OECD, 2019). Educators can effectively gauge students' mathematical proficiency by aligning the items with the desired assessment framework. Additionally, a comprehensive item development plan, including awareness of the time constraint in item formulation and provision of supplementary resources, all contribute to ensuring a systematic and organized approach to item writing. Moreover, Magno (2003) expressed that the level of appropriateness in test construction is operationally defined as the way in which item writers follow the essential guidelines, principles, and procedures in test construction.

Subtheme 1B: Enacting as the Climactic Transfiguration in Item-Writing

Suspense ensued during the participants' writing activities as they grappled with the challenge of what seemed like a puzzle game. It is evident, especially in the case of Ben, who said that it was a challenge for him to piece together the assigned content and context categories in a particular item and simultaneously satisfy them. Saanvi resonated with this concern as she prioritized "fitting" questions into the categories indicated by the predetermined TOS. She showed frustration whenever she deemed an item fit for the assigned content and context categories but not the assigned process category.

Another significant action point that emerged from the participants' narratives during the enactment phase was the use of the existing or released PISA items shown in the seminar as their basis for creating the contexts of their items. Ben said he constantly referred to the items in the slides shown during the seminar, which is the same case for Grace. Michaela sensed that including pictures in items was necessary, as suggested by the sample items in the seminar, and thus influenced her decision to do the same in her items. She backs up her action by explaining that pictures help supplement students' understanding of the question. In the case of Saanvi, she confessed that in creating items, she copied an existing item and then replaced the name of the character/s and created a story out of the item. Some effort was exerted in searching for suitable item resources such as pictures appropriate for the given item, and this was seen in Ben's ordeal in scouring the internet search engine results for a picture of a *Lumad* bracelet (*Lumad* – an indigenous ethnic group in Mindanao). He explicitly stated that he exerted considerable time and effort to find an appropriate photo that best depicts the Lumad bracelet he wanted to feature in his item. He did not want to settle for any other bracelet photo as he wanted to promote indigenous people's culture in his item, describing this deliberate choice as cultural sensitivity. Unlike Ben, who carefully considered his item context, Grace reportedly selected and randomly created item contexts.

The enacting phase involves the active creation of assessment items. In this phase, the most notable behavior among the respondents was typing the parts of each item until a completed structure became prominent. For multiple-choice items, such components are described by Baranowski (2006) as the stem, the lead line, and the response options, which can also be found in constructed-response items except for the last component. Additionally, it can be gleaned that the participants scrambled to consult various reference materials and gather different item resources, such as pictures, to supplement the item. Layouting also surfaced as a discernable action performed by the item writers. Most of their on-screen activities during this stage included adjusting the picture and table layout, size, and position.

The skills of mathematics educators related to mathematics content play a crucial role in this phase. Such knowledge has been labeled as mathematical content knowledge (MCK), which is known to support a mathematics teacher's pedagogical knowledge specific to mathematics (Tatto et al., 2020). The teachers' expertise ensures the items' accuracy, appropriateness, and rigor. To enhance the authenticity and relevance of the assessment, real-life contexts and problem-solving scenarios can be incorporated. By doing so, the student test-takers are challenged to apply their mathematical knowledge and skills to practical situations, reflecting the real-world value of mathematics.

Subtheme 1C: Reviewing as a Writer's Reflective Denouement

As each participant was given a time limit of four hours to create four items, they clamored that the given amount of time was insufficient, especially in reviewing the items. This concern was consistent across Grace's and Michaela's statements throughout the FGD. Michaela shared that she could review the first two items but not the last two items as she was already pressed for time, although it needed to be clarified how she reviewed them. There was also no concrete evidence from Ben on how he reviewed his items prior to submission, given that in the video recordings, he moved on to writing the next item just moments after finishing an item. His on-screen activities related to reviewing were mainly scrolling up and down the item, up to the start of the document, and down to the end of the document.

However, the expert's feedback on the participants' items allowed them to reflect on the items they created. As noted by Grace, she was generally satisfied with her items, except for one item where she felt she had a misconception of the scientific context, as the expert commented that one of her items had a context that was not scientific. Saanvi also realized that her questions were highly unrealistic, as the expert who reviewed her items pointed out. In effect, she adjusted some parts of her item contexts. Thanks to her post-training musings, she fortified her items with pictures to create some visual effect, allowing students to visualize the scenario presented in the problem.

The reviewing phase is critical in the item-writing process, as items are evaluated to ensure their quality and validity. On the writers' end, most of the reviewing activities on the screen were scrolling up and down the items, rereading the items' prompt, stimulus, and resources, reviewing correspondence of the item specifications with the TOS, and counterchecking the category definitions in the assessment framework. The last phase allowed the writers to revisit parts of their items that contained flaws that they initially overlooked. During the review process, the items' clarity, coherence, and fairness are assessed. Feedback from expert evaluators is also valuable in identifying any potential issues related to accuracy and clarity (Downing, 2006). Refinement based on feedback enhances the overall quality of the assessment items. At the same time, item review is subsumed under Step #4: Item Development which is part of Downing's (2006) twelve steps for test development.

Theme 2: Dimensions of Item-writing

The responses of the participants revealed that there were certain aspects to the activity of item-writing. These came in the form of cognitive, behavioral, and psychological dimensions. The cognitive dimension is primarily about the participants' thoughts, mental processes, and activities as they embark on the item-writing task. The behavioral dimension exposes the writers' idiosyncratic ways as they progress through the task. Finally, the psychological dimension describes the participants' emotions throughout the task.

Subtheme 2A – Cognitive Dimension: Filling the Cup

The item-writing seminar-workshop presented challenges and opportunities, as narrated by the participants. They were undoubtedly met with difficulties in thinking about the items, as in the words of Ben: "*It needs lots of brain cells for you to craft.*" The task challenged them mentally because they felt it required much mental effort, considering it was their first-time encountering PISA-like mathematics items. Ben cited that the mental strain was partly due to the struggle to satisfy all three categories simultaneously, referring to the content, context, and process categories. In particular, he described his dilemma where he felt that his question matched the assigned content category but would later encounter a problem with the assigned context category.

Grace echoed this same concern by saying that the content-context pairings posed a limitation that restricted their ideas and creativity. Unique to her case was the special mention of her trouble organizing ideas by arguing that teachers have many matters to attend to. She also believes that some teachers may have ideas but must be better equipped to handle many ideas as they are overwhelmed by their work. Grace also felt that she needed more resources as she only opted to refer to the training materials presented during the seminar when she could have referred to supplementary materials such as books and other references if she had enough time. Similarly, Saanvi shared her plight related to the resources to be used in her items as she expressed difficulty in finding pictures that were both appropriate and domain-free or those that are not protected by copyright and are freely available for use by the public.

As shown in their narratives, it was the participants' collective first experience in writing PISA-like mathematics items. Despite having some experience in writing test items for classroom use, they were somewhat naive to writing context-based mathematics items. The cognitive dimension of item writing is revealed by statements that pertained to the participants' thoughts as they traversed the writing task. The participants engaged in a problem-solving task in that they

tried to resolve arising conflicts in the predetermined content-context combination assigned to each item. Another strand under the cognitive dimension was the decision-making process, as the participants constantly faced a series of minor predicaments which were overcome by making small decisions, albeit on an impulse. The results corroborate Lasaten's (2016) finding that item writers considered authenticity, reliability, and validity of tests to be their greatest concerns. With regards their training needs, assessment of student learning in general came out to be the priority.

Subtheme 2B – Behavioral Dimension: Weaving the Fibers

The video recordings show that most participants follow a unidirectional smooth flow in creating the items. It means that each created the first item to completion before proceeding to the next, and this pattern was consistent until all four items were finished. However, Grace's experience stood out from the rest as she recounted her transition between items and described a nonlinear progression along the item-writing activity. For instance, she paused while writing the first item, moved to the next item, and then returned to the first item once she deciphered how to finish it. She also admits to occasionally getting stuck in writing an item but then jumps to another item in the interest of time and feels she is juggling between items simultaneously.

This vignette narrowly focuses on how writers have unique styles in accomplishing writing tasks. Such styles are influenced by individual factors such as experience, personality, preferences, and creativity. While unidirectional item-writing is a typical style that some writers share, this may not always be the case, as shown by one respondent. Patterns revealed by repetitive behaviors show that writing style may be one that an individual is the most comfortable with, primarily when he or she has found it to be helpful over time. Findings of the study echoed the results of Namoco and Zaharudin (2021) that the assessment of learning practices of the item writers are shaped by their pedagogical beliefs, social norms and intentions.

Subtheme 2C – Psychological Dimension: Traversing the Uncharted

The participants felt a range of emotions throughout the seminar-workshop, and this was prominent in their narratives. Ben, for one, felt mixed emotions when he learned that an expert on PISA would be commenting on his items. He even expressed that he is mere "dust" and is insignificant, comparing himself to the experts in terms of professional credentials. Both Ben and Grace expressed anxiety for different reasons: Ben's fear was induced by the daunting scope of the type of assessment involved, describing it as "large" and "international," while Grace could not get ahold of her ideas because of her belief that she was unorganized. Grace was also equally scared of receiving feedback from the expert evaluator as she feared that she may have made mistakes in her items and that she may have embarrassed herself.

Meanwhile, Grace and Michaela shared sentiments about the item-writing task conducted remotely online. Grace explained her situation at home during the seminar-workshop, being a mom tending to the needs of her family on a weekend. It is in the context of her concern regarding limited time in understanding the information from the seminar and planning for the items. Michaela, on the other hand, had her share of distractions at home, which she believes hindered her from concentrating on the training. It was because, during that time, she was taking care of her niece while participating in the study. She also voiced out that due to her circumstance, the quality of information she received during the seminar may have been compromised, which in turn may have affected the quality of her items. Finally, dismay and embarrassment were what Saanvi felt upon receiving feedback from her expert evaluator, who pointed out that one of her items had a highly unrealistic context. Table 1 below shows the expert's comments on one of Saanvi's items.

Table 1. Comments Given by an Expert to Saanvi's Item Q2

Saanvi's Item Q2	Expert's Comments
<i>Lany is a small girl. She usually can make 3 dresses out of 2-metre cloth. After 3 months since she broke up with her fiancé, she had stress eating and gained weight. And it results in a change in her fitting. Just a month ago, she started to go to the gym and do some exercise to go back to her size to attend her friend's wedding.</i>	"I may have a different perspective from yours, but I initially thought how a small girl (I linked it with age) could have a boyfriend and broke up with a boyfriend, be desperate and stressed of losing a boyfriend at a young age. So, a small girl is small because of her body size. What is the relevant SOCIAL CONTEXT here? What values are you communicating in this scenario? Just an opinion."
<i>If Lany used 30% of a 6-metre long cloth to make two blouses. How do we know that she gain or retained her size? Explain.</i>	"Be consistent on the tenses." [referring to the phrase "she gain or retained her size"]
Answer Key:	"How real is a girl, particularly with her looks and weight, to be wearing 3 the same dresses or blouses sewed from 2 meters of cloth."
<i>If Lany can have 3 dresses from 2 meters, we will divide 3 from 2m and we get 2/3 or 0.667m. After she had a month in the gym,</i>	

<i>she can now use 30% of 6m to make 2 blouses. That would give us 0.30×6 is 1.8m divided by 2 = 0.9. The answer is she gained weight and had a larger size compared to before.</i>	“The statements here, do not reflect “Formulate”. Have this align with the definition of Formulate. The problem though can be along “Formulate”
---	---

The second theme, *Dimensions of Item-writing* reveals various emotions that were pooled from the participants' narratives. These feelings can be grouped into two major clusters: positive (i.e., enlightenment, satisfaction) and negative (i.e., anxiety, embarrassment)- both natural human inclinations induced by a new, unfamiliar, and seemingly daunting task. Narratives on item-writing activities are anchored on emotions as these are intertwined with the participants' descriptions of their thoughts during the writing task. Unpacking the narratives shows evidence that cognitive and affective processes are interlocked, which explains why specific emotions can drive certain writing processes. This perspective is a crucial lens in trying to understand the participants' individual experiences in the writing task. This finding is supported by Rodriguez (1997) that item-writing requires an uncommon combination of special abilities. Each item, when it is written, introduces new difficulties and opportunities. There can be no fixed formulas for making a good story, just as there can be no established standards for producing good test items. It is the item writer's personality, including emotions, that shapes the crafted items.

Discussion

The findings emphasize three important phases in item-writing: planning, enacting, and reviewing. These results are aligned with the cognitive model of item-writing by Fulkerson et al. (2009, as cited in Fulkerson et al., 2010), where the representation phase is parallel to the planning stage in the present study, the exploration phase is observed in the enacting part, and the solution phase is likened to the reviewing portion. The emerging themes of the present study clearly support the existing body of knowledge.

The participants admitted that they had difficulties in jumpstarting their item writing; these were evident as they scrambled through references such as the PISA assessment framework and released PISA math items. They felt that such difficulties would have been alleviated if they were given sufficient time to organize their thoughts, map and iron out their plans. This goes to show how planning is an important phase in item-writing, particularly setting objectives and a well-defined purpose for assessment. Similar challenges were encountered by Serbian teachers in Radišić and Baucal's (2018) study, as resonated in their reflections about PISA items. Most of them have difficulty communicating which procedures the students are expected to follow to be able to solve the mathematical tasks, while disagreement among the teachers is noticeable on identifying the most difficult part of each item (Radišić & Baucal, 2018).

In the case of the participants' task, understanding the different types of content domain categories (content categories, context categories and process categories) seemed to be the crucial element. As the participants enacted item-writing, they were fleshing out their thoughts and ideas albeit on the fly. This is evidenced by their formulation of contexts which are rooted in their backgrounds and experiences. It became very helpful in item-writing because contexts are story-like texts that require building up and hence take up time. By banking on their personal and subjective knowledge, they were able to generate ideas for their items instantaneously. This important observation on the results was manifested by Memisevic and Biscevic (2022) on the weight of culture-related factors in PISA. Also, the banking on personal experience in framing the test items context was supported by Kohar et al. (2014) as they found that the prototype items were linked to personal experiences.

Additionally, most of the mathematics teachers spent a significant amount of time performing item editing and layouting, including item resources such as tables and images. Finally, reviewing afforded the participants a chance to revisit their items. If not immediately after the item-writing seminar-workshop, at least they had this chance after receiving the comments of the expert evaluators. External feedback made them realize some aspects of their item that needed improvement. Those who were able to review their items engaged in behaviors like rereading the problem and making minor edits in both text composition and layout. Consequently, the challenges encountered by the participants corroborated in the study of Zulkardi and Kohar (2018) that authenticity of the context, accessibility of the language used, and the demand for higher-order thinking skills are validated in the results of the study.

Not only do the findings shed light on the phases of item-writing, but it also brought three dimensions of item-writing to the spotlight. The results from this study showed that item-writing activities can be viewed through a multimodal perspective, involving three concurrent aspects: the cognitive dimension, behavioral dimension and psychological dimension. As Fulkerson et al. (2009, as cited in Fulkerson et al., 2011) and other scholars viewed item-writing as cognitive in nature, the research results introduce the idea that the interplay between cognitive, behavioral, and psychological dimensions form part of the item-writing phenomenon. Item-writing is like an Olympiad performance that cannot be explained singularly but, in its plurality, as Bicar and Gaylo (2022) described it. The multi-dimensional feature of item-writing phenomenon as documented in the present paper contributes to the fund of knowledge existing in the item-writing literature. It implies a new trajectory on dealing with item-writing competence in the education field. Cognitive, behavioral, and psychological areas on item-writing need to be considered in enhancing the skills of professionals associated with writing test items.

For the cognitive dimension, the participants expressed mental distress in their partaking of the task, which is a highlight in their collective sentiments. Their behaviors in the task showcased that the participants had unique styles in writing which were influenced by individual factors, as seen in their varied approaches in crafting the contexts for their items. With regards to the psychological dimension, the participants were understandably intimidated as they found the task daunting, leaving them to wonder about how to construct their items. Such observations were to be expected given that all four participants have never experienced being subjected to the task of writing PISA-like math items. Although the PISA mathematics assessment framework was discussed a few hours before the item-writing workshop, it was their first time encountering the PISA mathematics assessment framework and reading its fine print up close.

Conclusion

Writing PISA-like mathematics items is a multifaceted endeavor encompassing several phases and dimensions. The planning, enacting, and reviewing phases provide a structured approach to item development, ensuring the quality and validity of the assessment. Meanwhile, the cognitive, behavioral, and psychological dimensions highlight the teachers' personal experiences and the conscious nature of item writing. These dimensions reflect the reality that the teachers face as they go through item-writing which further enhances the comprehensiveness and authenticity of the items. Considering these elements adds more depth to understanding the phenomenon of item-writing, something that research in mathematics education gives inadequate attention to. Recognizing the multifaceted nature of writing PISA-like mathematics items and considering the phases and dimensions involved have significant implications. For one, developing high-quality assessment items that effectively measure students' mathematical abilities and yield valid and reliable results is a complex process laced with intricacies at the writer's level. Paying attention to this phenomenon can allow educators and assessment developers to take a closer look at how to better support writers as they embark on this task.

This consequently leads to an improved assessment quality, a better understanding of students' mathematical abilities, and more authentic assessment practices. All of which, in turn, contribute to the slow but continuous improvement in the field of mathematics education and foster global benchmarking among educational institutions. Implementing the planning, enacting, and reviewing phases allows for careful item development, leading to higher-quality assessments. Such, in turn, ensures that the assessments accurately measure students' mathematical abilities, providing more reliable and valid results, and these influence mathematics teaching and learning over time.

Recommendations

The researchers recommend that similar studies in the future look into the quality of the PISA-like items constructed by the teachers. Additionally, this can be done by developing a framework anchored on both a theoretical basis (e.g., evaluating difficulty levels of mathematical items based on a taxonomy) and on the PISA mathematics assessment framework. Furthermore, future researchers might want to replicate the study but observe the item-writing activities of mathematics teachers in basic education (e.g., elementary and high school mathematics teachers), given that ILSAs such as PISA assess students in basic education.

Consequently, mathematics educators may utilize PISA-like mathematics items in their classroom assessments to further enhance the mathematical proficiency of the students, wherein the items are aligned with international standards. School administrators may initiate capacity building activities to enable mathematics teachers to competently produce PISA-like mathematics items, considering the three dimensions uncovered in the study. Concerned parents may also support the initiatives of the teachers and administrators to be at par with high-performing schools, both locally and internationally, on international large-scale assessments like PISA.

Limitations

The item-writing activities that were observed and the experiences of the participants were analyzed under the set-up where they constructed items individually, without any interaction with the rest of the participants.

Ethics Statement

This study which involved human participants was reviewed and approved by the Research Ethics Committee of the Ateneo de Manila University (Protocol ID: SOSEREC_22_001). The participants provided their written informed consent to participate in this study.

Acknowledgement

The authors would like to express their gratitude to Bukidnon State University for the generous assistance extended and for granting the primary author the permission to conduct the pilot study of his dissertation project.

Conflict of Interest

The authors have no conflicts of interest to declare.

Funding

This research was made possible by the Office of the Assistant Vice President for Graduate Education (formerly Office of the Associate Dean for Graduate Programs) of the Ateneo de Manila University which provided the primary author a scholarship grant for his studies.

Authorship Contribution Statement

Garcia: Conceptualization, data analysis, drafting manuscript. Gaylo: Data analysis, data interpretation, organization of themes. Vistro-Yu: Reviewing, supervision, final approval.

References

- Al-Bahlani, S. M. (2019). *Assessment literacy: A study of EFL teachers' assessment knowledge, perspectives, and classroom behaviors* [Doctoral dissertation, The University of Arizona]. University of Arizona Repository. <http://hdl.handle.net/10150/633240>
- Baranowski, R. A. (2006). Item editing and editorial review. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 349-357). Routledge. <https://doi.org/10.4324/9780203874776>
- Bernardo, A. B. I. (2021). Socioeconomic status moderates the relationship between growth mindset and learning in math and science: Evidence from PISA 2018 Philippine data. *International Journal of School and Educational Psychology*, 9(2), 208-222. <https://doi.org/10.1080/21683603.2020.1832635>
- Bicar, V., & Gaylo, D. (2022). Cluster characterization of countries' performance in mathematics olympiad: Input to mathematics education. *Science International (Lahore)*, 34(5), 503-506. <https://doi.org/10.5281/zenodo.7240929>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp0630a>
- Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57-71). American Psychological Association. <https://doi.org/10.1037/13620-004>
- Cagasan, L., Luo, R., Robertson, P., & Care, E. (2016). *Formative assessment project: Phase 2 research report*. Assessment, Curriculum, and Technology Research Centre (ACTRC). Melbourne and Manila. <http://bit.ly/3up0Ej3>
- Caine, V., Clandinin, D. J., & Lessard, S. (2022). *Narrative inquiry: Philosophical roots*. Bloomsbury Publishing. <https://doi.org/10.5040/9781350142084>
- Chi, C. (2023, December 6). *Philippines still lags behind world in math, reading and science — PISA 2022*. Philstar Global Corp. <https://bit.ly/3whiHLx>
- Clores, M. A., & Reganit, A. A. R. (2020). Investigating the assessment literacy of teachers in private junior high schools in the Philippines. *Humanities, Arts and Social Sciences Studies*, 20(2), 461-476. <https://doi.org/10.14456/hass.2020.17>
- Close, S., & Shiel, G. (2014, June 24-25). *A Comparison of TIMSS 2011 and PISA 2012 math frameworks and performance for Ireland and selected countries* [Paper Presentation]. Science and Mathematics Education Conference, Dublin City University.
- Department of Education. (2019). *PISA 2018: National report of the Philippines*. <https://bit.ly/3SYYSbq>
- Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Lawrence Erlbaum.
- Duff, P. A. (2012). How to carry out case study research. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 95-116). Blackwell Publishing.
- Elshawa, N., Heng, C. S., Abdullah, A. N., & Rashid, S. (2016). Teachers' assessment literacy and washback effect of assessment. *International Journal of Applied Linguistics and English Literature*, 5(4), 135-141. <https://bit.ly/3UlhCri>
- Espinosa, A. A., Gomez, M. A. C., Reyes, A. S., Macahilig, H. B., Cortez, L. A. S., & David, A. P. (2023). International large-scale assessment (ILSA): Implications for pre-service teacher education in the Philippines. *Issues in Educational Research*, 33(2), 553-569. <https://www.iier.org.au/iier33/espinosa.pdf>
- Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, 21(4), 357-364. <https://doi.org/10.1016/j.tate.2005.01.008>

- Fulkerson, D., Nichols, P., & Mittelholtz, D. (2010, May 3). *What item writers think when writing items: Towards a theory of item-writing expertise* [Paper Presentation]. Annual Meeting of the American Educational Research Association, Denver, CO.
- Fulkerson, D., Nichols, P., & Snow, E. (2011, April 8-12). *Expanding the model of item-writing expertise: Cognitive processes and requisite knowledge structures* [Paper Presentation]. Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Gaylo, D. N., Caingcoy, M. E., & Mugot, D. C. (2020). Utilization of scholarly journal articles in the teaching and learning of teacher education courses. *Balkan and Near Eastern Journal of Social Sciences*, 6(3), 59-66. <https://doi.org/10.47696/adved.202038>
- Geisinger, K. F., & Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement from foundations to future* (pp. 3-20). The Guilford Press.
- Griffin, P., Cagasan, L., Care, E., Vista, A., & Nava, F. (2016). Formative assessment policy and its enactment in the Philippines. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 75-92). Springer. https://doi.org/10.1007/978-3-319-39211-0_5
- Hailaya, W. M. (2014). *Teacher assessment literacy and student outcomes in the Province of Tawi-Tawi, Philippines* [Doctoral thesis, The University of Adelaide]. <https://bit.ly/3I0BeLD>
- Hanafi, N. B. M., Ali, N. B. M., Zamani, S. B., Yamin, N. A. B., & Ismail, N. N. B. (2020). Examining assessment literacy: A study of technical teacher. *European Journal of Molecular and Clinical Medicine*, 7(8), 705-717. <https://bit.ly/311OT4Y>
- Haw, J. Y., King, R. B., & Trinidad, J. E. R. (2021). Need supportive teaching is associated with greater reading achievement: What the Philippines can learn from PISA 2018. *International Journal of Educational Research*, 110, Article 101864. <https://doi.org/10.1016/j.ijer.2021.101864>
- International Association for the Evaluation of Educational Achievement. (2020). *TIMSS 2019 trends in international math and science study Philippines country report*. <https://bit.ly/4bG4DbM>
- Johnson, M., Constantinou, F., & Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal*, 43(4), 700-719. <https://doi.org/10.1002/berj.3281>
- Khalid, N. H. M., Latif, A. A., & Yusof, I. J. (2021). Assessment literacy: A systematic literature review and research agenda. *Annals of the Romanian Society for Cell Biology*, 25(3), 4668-4696. <https://bit.ly/3uG31hd>
- King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research* (pp. 256-270). Sage. <https://doi.org/10.4135/9781446280119.n21>
- Kohar, A. W., Zulkardi, Z., & Darmawijoyo, D. (2014). Developing PISA-like math tasks to promote students' mathematical literacy. In R. Ilma (Ed.), *Proceeding in the Second South East Asia Design - Development Research (SEA-DR) Conference* (pp. 14-26). Universitas Sriwijaya.
- Kuger, S., & Klieme, E. (2016). Dimensions of context assessment. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning* (pp. 3-37). Methodology of Educational Measurement and Assessment. Springer. https://doi.org/10.1007/978-3-319-45357-6_1
- Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, 36, 143-152. <https://doi.org/10.1016/j.tate.2013.07.010>
- Lapinid, M. R. C., Cordell, M. O., II, Teves, J. M., Yap, S. A., Chua, U., & Bernardo, A. B. I. (2022). *Addressing the poor mathematics performance of Filipino learners: Beyond curricular and instructional interventions*. *DLSU-AKI Policy Brief*, 14(2), 1-4. <https://bit.ly/3WmQ9lp>
- Lasaten, R. C. S. (2016). Assessment methods, problems and training needs of public high school teachers in English. *International Journal of Languages, Literature and Linguistics*, 2(2), 55-60. <https://doi.org/10.18178/ijll.2016.2.2.67>
- Lian, L. H., Yew, W. T., & Meng, C. C. (2014). Enhancing Malaysian teachers' assessment literacy. *International Education Studies*, 7(10), 74-81. <https://doi.org/10.5539/ies.v7n10p74>
- Magno, C. (2003). The profile of teacher-made test construction of the professors of University of Perpetual Help Laguna. *UPHL Institutional Journal*, 1(1), 48-55. <https://ssrn.com/abstract=1429347>
- Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education*, 43(6), Article 1. <https://doi.org/10.14221/ajte.2018v43n6.1>

- Memisevic, H., & Biscevic, I. (2022). Mathematics, gender and the meaning in life: The results of PISA testing in Bosnia and Herzegovina. *European Journal of Mathematics and Science Education*, 3(2), 171-179. <https://doi.org/10.12973/ejmse.3.2.171>
- Montemayor, M. T. (2023, December 6). *CHED to address PH students' low int'l assessment ranking*. Philippine News Agency. <https://www.pna.gov.ph/articles/1215002>
- Mullis, I. V. S., Martin, M. O., & Loveless, T. (2016). *20 years of TIMSS: International trends in math and science achievement, curriculum, and instruction*. International Association for the Evaluation of Educational Achievement (IEA). <https://bit.ly/3UESGN8>
- Namoco, S., & Zaharudin, R. (2021). Pedagogical beliefs and learning assessment in Science: Teacher's experiences anchored on theory of reasoned action. *Journal of Turkish Science Education*, 18(2), 304-319. <https://doi.org/10.36681/tused.2021.67>
- Napanoy, J. B., & Peckley, M. K. (2020). Assessment literacy of public elementary school teachers in the indigenous communities in Northern Philippines. *Universal Journal of Educational Research*, 8(11B), 5693-5703. <https://doi.org/10.13189/ujer.2020.082203>
- Orbeta, A. C., Melad, K. A. M., & Potestad, M. (2020). *Correlates of test performance of 15-year-old students in the Philippines: Evidence from PISA* (No. 2020-57). PIDS Discussion Paper Series. <https://bit.ly/3T03sw8>
- Organisation for Economic Cooperation and Development. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. <https://bit.ly/44KfcC>
- Piosang, T. L. (2017). A cross-sectional analysis of classroom assessment literacy of English teachers in secondary and tertiary levels. *Educational Measurement and Evaluation Review*, 8(1), 30-48. <https://bit.ly/3y0d2qX>
- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, 46(4), 265-273. <https://doi.org/10.1080/08878730.2011.605048>
- Radišić, J., & Baucal, A. (2018). Teachers' reflection on PISA items and why they are so hard for students in Serbia. *European Journal of Psychology of Education*, 33, 445-466. <https://doi.org/10.1007/s10212-018-0366-0>
- Rodriguez, M. C. (1997, March 24-28). *The art & science of item-writing: A meta-analysis of multiple-choice item format effects* [Paper Presentation]. Annual Meeting of the American Educational Research Association, Chicago.
- Rodriguez, M. C., & Haladyna, T. M. (2013). Writing selected-response items for classroom assessment. In J. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 293-312). SAGE Publications, Inc. <https://doi.org/10.4135/9781452218649.n17>
- Schleicher, A. (2019). *PISA 2018: Insights and interpretations*. OECD Publishing. <https://bit.ly/4bG36K>
- Sonday, A., Ramugondo, E., & Kathard, H. (2020). Case study and narrative inquiry as merged methodologies: A critical narrative perspective. *International Journal of Qualitative Methods*, 19, 1-5. <https://doi.org/10.1177/1609406920937880>
- Tatto, M. T., Rodriguez, M. C., Reckase, M. D., Smith, W. M., Bankov, K., & Pippin, J. (2020). The FIRSTMATH study: Concepts, methods, and strategies for comparative international research in mathematics education. In M. T. Tatto, M. C. Rodriguez, M. D. Reckase, W. M. Smith, K. Bankov, & J. Pippin, *The first five years of teaching mathematics (FIRSTMATH): Concepts, methods, and strategies for comparative international research in math education* (pp. 1-20). Springer International Publishing. https://doi.org/10.1007/978-3-030-44047-3_1
- United Nations Children's Fund. (2021). *SEA-PLM 2019 Southeast Asia primary learning metrics Philippines country report*. <https://bit.ly/49cBIQ5>
- United Nations Children's Fund & Southeast Asian Ministers of Education Organization. (2019). *SEA-PLM 2019 Assessment Framework* (1st ed.). United Nations Children's Fund (UNICEF) & Southeast Asian Ministers of Education Organization (SEAMEO) – SEA-PLM Secretariat. <https://bit.ly/3SWISOA>
- Wu, M. L. (2009, June 2). *A critical comparison of the contents of PISA and TIMSS mathematics assessments* [Paper Presentation]. NCES "What we can learn from PISA" research conference, Washington, DC.
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Zulkardi, Z., & Kohar, A. W. (2018). Designing PISA-like math tasks in Indonesia: Experiences and challenges. *Journal of Physics: Conference Series*, 947, Article 012015. <https://doi.org/10.1088/1742-6596/947/1/012015>