

**VALIDATION OF BILINGUAL PISA MATHEMATICS ITEMS: A
COMPARATIVE STUDY OF ENGLISH AND FILIPINO VERSIONS IN
A PHILIPPINE PUBLIC SCHOOL**

**¹Freddielyn B. Pontemayor, ²Cherly C. Cordova, ³Jenyliza T. Uchang,
⁴Jessel T. Torres, ⁵Denis A. Tan, ⁶Rhoderick V. Nuncio,**

^{1,2,3,4,5}College of Education, Central Mindanao University, Bukidnon, Mindanao, Philippines.

⁶Full Professor, De La Salle University

Abstract

The study aimed to validate the original and Filipino-translated versions of the PISA math test items as preparatory instruments for a quasi-experimental research design involving Philippine secondary school students. Seventy-three Grade 9 students from two intact classes at Dologon National High School were included in the January 2025 pilot testing. One group had to respond to the English version of the test, while the other group was given the Filipino-translated version. The instrument, which consisted of 37 items and was designed to align with the PISA math framework, was administered within a 1.5-hour time limit and subjected to reliability and item-level analyses.

The English translation had a Cronbach's alpha coefficient of 0.707, which is an acceptable internal consistency based on standard psychometric criteria (George & Mallery, 2003). The Filipino translation, however, initially had a relatively marginal alpha of 0.639, which, while less than optimal, reflects the usual challenges with translated and culturally adapted tests (Hambleton & Patsula, 1999). By the removal of items with zero or negative item-total correlations, the Filipino version then had its reliability substantially enhanced to 0.739 for the remaining 30 items and thus satisfied the minimum reliability threshold for exploratory studies in educational measurement.

This study offers empirical support for the internal consistency of both versions, suggesting that language translation—when thoroughly examined and empirically tested—is not necessarily at the cost of the integrity of cognitive measurement. This study provides empirical evidence supporting the internal consistency of both versions, indicating that language translation—when rigorously evaluated and empirically validated—does not inherently compromise the integrity of cognitive measurement. After item revision, the Filipino version showed not only greater reliability but also greater concordance between item content and the underlying mathematical problem-solving construct. This study emphasizes the value of item-level diagnostics in the validation of bilingual tools, particularly in multilingual environments where students may have different degrees of proficiency in the language of assessment. The study also highlights the need for the adaptation of international

assessment frameworks like PISA to enable fair measurement and policy-relevant comparisons across diverse groups of learners.

Keywords: *PISA Mathematics, Instrument Validation, Bilingual Assessment, Filipino Translation, Language and Assessment*

Introduction

The Programme for International Student Assessment (PISA), administered by the Organization for Economic Co-operation and Development (OECD), functions as a globally recognized benchmark for evaluating and comparing the effectiveness of educational systems. It assesses the competencies of 15-year-old students in reading, mathematics, and science, providing data-driven insights into students' preparedness for real-world challenges. Given the fact that PISA aims to measure the extent of students' capability of using knowledge in real-world contexts, its findings have become essential measures of educational quality as well as policy drivers at the international level. Nevertheless, cross-national comparison of such measures depends on the assumption that the measurement tools operate equally well in both languages and cultures—an assumption that proves particularly challenging in multilingual environments.

In countries like the Philippines, English and Filipino are both official languages, and a third regional dialect is widely spoken by students in the home environment. This language orientation of Filipino learners contributes to the uniqueness of its learners. It is sometimes contributory to either positive or negative performances of students in tests. The test language has a direct effect on students' performances as studied by Bernhofer and Tonin (2022), which found that taking an exam in a second language results in a significant loss in grade points—approximately 9.5%—indicating a direct negative effect of test language on student performance. More and more research indicates how linguistic factors affect not only the comprehension of test questions by students but also how they can communicate higher-order thinking skills. Even though English is employed as a medium of instruction in most schools in the Philippines for mathematics, variations in students' levels of English proficiency create construct validity and fairness of measurement issues when English-only tests are administered. *J Intell.* (2022) highlights that language abilities significantly influence cognitive test performance, emphasizing that examinees' expressive and receptive language skills affect their test outcomes more than test characteristics themselves.

This research examines a critical deficit in the validation of assessment through a psychometric examination of a bilingual form of the PISA mathematics test, in English and Filipino. Grounded in the theoretical underpinnings of Classical Test Theory (CTT) and the related methodologies of instrument adaptation, the research aims to determine if the two language forms of the test produce data that is consistently internal and comparable. Prepiloting was initially carried out at Dologon National High School among Grade 9 students, within a larger initiative to create valid and equitable assessment instruments for application in a quasi-experimental research design.

By comparing item-specific performance and internal reliability of the two versions, the study aims to establish the type of linguistic or cultural differences translation generates and evaluate their influence on the psychometric quality of the test. The main aim is to establish whether the translated version maintains conceptual and functional equivalence with the original version, thereby making its use in subsequent studies of language influence on mathematical attainment possible. The findings are intended to inform national initiatives to calibrate international large-scale assessments to enhance the equity, inclusiveness, and diagnostic quality of standardized tests in diverse education contexts.

Review of Related Literature and Conceptual Framework

PISA and Terminology of Evaluation

The Programme for International Student Assessment (PISA) has become an international standard for measuring education systems based on how students' abilities to use their knowledge in real life. However, one of the ongoing problems of cross-national testing deals with the effect of the assessment language on student performance. Although PISA provides detailed frameworks and standard items, the assessment is carried out in different languages to suit the participating countries' needs. The requirement raises one of the central psychometric questions: Can a translated assessment measure the same construct with the same degree of accuracy as the original assessment?

Linguistic variation has been shown to affect students' performance, particularly in multilingual nations where the instruction language is not necessarily the students' home language. Studies have all shown that when tests are given in a second language or in a language in which students have weaker proficiency, the measured performance would reflect language deficiencies and not true mathematical reasoning capacity (Abedi, 2002; SolanoFlores & Trumbull, 2003). However, the relative contributions of specific features of linguistic complexity to this disadvantage are largely unclear (Haag N., Heppt B., Stanat P., Kuhl P., Pant H. 2013) while Henry, Nistor, & Baltes (2014) found that English proficiency predicts Mathematics performance which point to language proficiency as one of the factors affecting performance which is supported by Peng et al. (2020). In the Philippine case, where English is most often used in instruction but Filipino is widely used at home, linguistic accessibility of test materials is a serious issue of equity.

Translation and validation of tools in multi-lingual settings.

Test translation is not an easy language task, but a complex psychometric task that requires careful attention to semantic, idiomatic, experiential, and conceptual equivalence (Hambleton & Patsula, 1999). A poorly translated test item can lead to differential item functioning (DIF), whereby equivalent items may behave differently in different language versions but measure the same construct. Therefore, stringent validation procedures are required to ascertain that the translated items maintain construct validity and measurement invariance across groups.

Recent studies emphasize the importance of adaptation fidelity and translation validity in cross-national large-scale assessments (ILA) in diverse cultural contexts (OECD, 2019;

Ercikan & Lyons-Thomas, 2013). Methods such as back-translation, committee review, and expert panels are widely used to minimize bias. Still, these methods must be supplemented with empirical validation through field testing and psychometric examination. Methods such as Cronbach's alpha, corrected item-total correlation, and exploratory factor analysis are used on an ongoing basis to assess internal consistency and item quality in translated testing.

Conceptual Framework: Classical Test Theory (CTT) as Validation Framework

This research is grounded on the assumptions of Classical Test Theory (CTT) as a fundamental model in educational measurement. Under CTT, an observed test score can be partitioned into a true score and an associated error component. CTT yields crucial measures for ascertaining the reliability and internal structure of tests, including measures like Cronbach's alpha as a measure of internal consistency and item-total correlation coefficients as test measures for item discrimination. Its primary aim is to evaluate and maximize the reliability and validity of test scores by minimizing measurement error. CTT can be used to analyze item-level statistics, specifically item discrimination, as well as overall test reliability and standard error of measurement, providing a comprehensive approach to test validation.

In this research, CTT is applied as a model to evaluate the psychometric quality of the original English and Filipino translations of PISA mathematics items. Through item behavior and internal consistency analysis of the two sets, the research seeks to establish whether the Filipino translation possesses equal construct equivalence and psychometric integrity as the original test. The model enables data-driven decisions for item retention, deletion, or revision, thus rendering the instruments more reliable and valid to apply in bilingual or multilingual educational contexts.

Methodology

Research Design

This study employed a quantitative and descriptive psychometric theory-driven research design with a specific focus on instrument validation. The main objective was to determine both the internal consistency and item-level statistics of the individual items of the bilingual versions—English and Filipino—of a mathematics test based on the Programme for International Student Assessment (PISA) framework. This research design was employed to present empirical evidence regarding the reliability and structural soundness of the instruments before they were employed in a quasi-experimental intervention.

Participants

The pilot administration was done in January 2025 with Dologon National High School, a Philippine public high school. Two intact Grade 9 classes with a total of 73 students in two homogeneous classes were used. Students in one class were given the English form of the PISA mathematics items, and the other class was given the Filipino translation. Using naturally occurring class groups ensured ecological validity and replicated the actual classroom settings where the instruments will be administered in the larger quasi-experimental study.

Instruments

The test instrument was 37 adapted mathematics items drawn from publicly accessible PISA assessment materials. The selection was made to span a range of cognitive levels and subject domains (e.g., quantity, space and shape, and change and relationships). The English version was the base instrument. For the Filipino version, a committee translation approach was employed, using mathematics educators and bilingual language specialists. This approach provided linguistic precision, conceptual equivalence, and cultural adaptability in the translation items. Both versions were piloted in timed, standardized testing conditions, with students completing the test in 1.5 hours.

Data Analysis Procedures

Data were examined with Classical Test Theory (CTT) methods to determine internal consistency and item discrimination. Namely, Cronbach's Alpha was computed for both versions to measure the internal consistency reliability of the instrument. A coefficient alpha of ≥ 0.70 was sufficient for the research instruments at a group level (Nunnally & Bernstein, 1994). Corrected Item-Total Correlations (CITC) were computed to identify the degree to which each item correlated with the total test score, excluding itself. Items with CITC coefficients of less than 0.20 were reported to be considered for revision or deletion (DeVellis, 2017). Also, alpha if Item Deleted to determine whether deleting specific items would improve overall reliability. Items with zero, negative, or low discrimination indices were deleted systematically, and the reliability of the revised instrument was retested to assess improvements in psychometric quality.

Results and Discussion

The pilot test results show some clear differences in how the original English and the translated Filipino versions of the PISA math assessment perform. The initial Cronbach's alpha for the English version was 0.707, which is considered acceptable for educational tests (George & Mallery, 2003; Nunnally & Bernstein, 1994). This indicates that the English version tends to measure math skills reliably, so it can be confidently used in research and classroom settings.

On the other hand, the Filipino version started with a Cronbach's alpha of 0.639, which is just below the acceptable range. While it is close, this suggests there are some inconsistencies in how the test items work together—likely because of translation issues or cultural differences. Looking closely at individual items, several—specifically Items 2, 5, 7, 19, 27 to 29, and 32 to 37—showed negative or zero correlations, meaning they didn't really add to measuring the main idea. This matches previous studies that emphasize how translation errors, cultural mismatches, and idioms can threaten the validity of assessments across languages (Hambleton & Patsula, 1999; Ercikan & Lyons-Thomas, 2013).

After removing the problematic items based on standard test theory guidelines, the Filipino version was shortened to 30 questions. Recalculating the reliability, the new Cronbach's alpha went up to 0.739. This improvement shows that carefully selecting and refining items helps

make the test more precise. More importantly, it confirms that when adapting assessments for different languages and cultures, you need to combine careful translation with thorough psychometric testing to ensure the test still measures what it is supposed to measure (Van de Vijver & Hambleton, 1996).

The findings implied that, first, they emphasize that language is not just a neutral tool—it really influences test results. In countries like the Philippines, where math lessons often happen in English, but many students speak Filipino or regional dialects at home, students might struggle more with understanding the language of the test than with the math itself (Abedi, 2002; Solano-Flores & Trumbull, 2003), Peng et al. (2020), Henry, Nistor, & Baltes (2014). These studies support the idea that language challenges can make it harder for students to show their true math skills, which might hide their actual academic ability.

Second, the results emphasize the importance of creating fair, culturally sensitive testing tools both locally and internationally. Improving the Filipino version not only made it more reliable but also helped develop a tool that respects the linguistic and cultural context of Filipino students. Culture plays a significant role in psychometric testing, affecting how individuals' approach and respond to assessment tools. Research has shown that cultural differences can impact test performance, with some studies indicating that cultural background can influence scores on cognitive abilities tests by as much as 15 to 20 points (Herrnstein & Murray, 1994). This is especially important for countries trying to balance global assessments like PISA with local educational needs. Success in these efforts depends on careful, repeated steps of translation, adaptation, and validation, rather than just translating words directly.

Third, this study adds to the ongoing conversation about fairness and equity in educational testing. When students take assessments in a language that doesn't quite match their cognitive and language backgrounds, it can lead to unfair results or misinterpretations of what they really know. The improved Filipino version, now stronger in its measurement qualities, offers a good alternative especially in situations where English skills might get in the way, making assessments fairer and more accessible.

Finally, having reliable test tools in both English and Filipino gives us a solid basis for future research. These tools make it easier to explore how language influences math problem solving, because we can compare groups using tools that are equally valid in each language. The methods used in this study also provide a model that other multilingual countries can follow when adapting international assessments for their local languages.

Overall, these findings emphasize how language, culture, and thinking skills are all connected in large-scale tests. They also point to practical ways to make assessments fairer and more reliable in bilingual settings. This supports the larger goal of creating testing systems that are fair and respectful of learners' diverse backgrounds, especially as education becomes more global but still rooted in local contexts.

CONCLUSIONS

The study assess the psychometric soundness of the English and Filipino versions of the modified PISA math items. The results showed that both versions had acceptable internal consistency, with the Filipino version improving after item refinement and the English version yielding an acceptable Cronbach's alpha . These findings demonstrate that the construct validity and reliability of assessment instruments across languages can be preserved through precise translation, cultural adaptation, and empirical validation.

The findings highlight the importance of language in influencing students' test scores, especially in multilingual settings like the Philippines, where Filipino is still the primary language of communication despite English being the medium of instruction. The updated Filipino version demonstrates how assessments that are culturally sensitive and linguistically accessible can yield more equitable and representative measurements of student aptitude. The continuous quest for inclusivity and equity in educational testing is aided by these findings.

The study has significant implications for classroom instruction and educational policy in addition to psychometric validation. Validated bilingual tools can be used by education organizations like the National Education Testing and Research Center (NETRC) and the Department of Education (DepEd) to create more equitable national tests and get students ready for global, extensive assessments like the Programme for International Student Assessment (PISA). Instead, teachers can use these modified tools to diagnose learning gaps, guide instruction, and improve mathematical literacy in both Filipino and English contexts.

Also, this study reaffirms that validation and translation are crucial steps toward educational equity rather than just being linguistic exercises. Teachers and legislators can encourage more accurate assessment of students' cognitive abilities by making sure that tests are both psychometrically sound and culturally sensitive, laying a more equitable basis for both domestic benchmarking and cross-border comparison.

Recommendations

Here are some suggestions to help improve the development of reliable and valid bilingual assessment tools.

First, future research should include both exploratory and confirmatory factor analysis (EFA/CFA) to understand how well the test items measure the core concepts of math literacy across different language versions. This will give a clearer picture of whether the questions are truly capturing what they're supposed to, and if the structure holds up across languages.

Second, it's important to increase the number and diversity of participants in these studies. Testing the instruments across various regions, social backgrounds, type of institution (private or public) and languages will help ensure that the results are more universally applicable and that the tool works well for all kinds of students.

Third, researchers may consider methods like cognitive interviews—such as think-aloud protocols or asking students to explain their reasoning—to see how students understand and

respond to translated questions. These qualitative approaches can reveal issues like language ambiguities or cultural mismatches that numbers alone cannot capture.

Fourth, once the assessment items, especially those aligned with global standards like PISA, are thoroughly validated, they should be made openly available. Sharing these tools can support teachers and schools in planning lessons, diagnosing student skills, and supporting further research, all while respecting local languages and contexts. The poor performance of Filipino students is a nationwide concern and the best way to address this problem is by coming up with a unified effort like this.

Finally, education policymakers need to incorporate these tools into national assessments. The Department of Education and related agencies should consider using validated bilingual assessments alongside existing tests like the NAT as tools to gauge readiness for international exams. This can help drive reforms, create more comprehensive policies, and better reflect the multilingual reality of Filipino learners.

References

1. Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, 8(3), 231–257. https://doi.org/10.1207/S15326977EA0803_02
2. DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.
3. Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in multiple languages and cultures. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 623–648). American Psychological Association. <https://doi.org/10.1037/14049-035>
4. George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (11.0 update, 4th ed.). Allyn & Bacon.
5. Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, 1(1), 1–30. <https://www.testpublishers.org/assets/jatt/hambleton.pdf>
6. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
7. Organisation for Economic Co-operation and Development (OECD). (2019). *PISA 2018 assessment and analytical framework: Mathematics, reading, science and global competence*. OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
8. Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13. <https://doi.org/10.3102/0013189X032002003>
9. Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1(2), 89–99. <https://doi.org/10.1027/10169040.1.2.89>

10. Nugky Permatasari, A., Firmansyah, F., & Retnawati, H. (2024). The effect of test preparation on English proficiency performance of English learners: A meta-analysis. *International Journal of Language Testing*, 14(2), 104–116. <https://doi.org/10.22034/IJLT.2024.432975.1316>
11. Maarouf, S. (2024). Examining the Effects of Language Competencies on Academic Achievements of Special Needs Students and Their Peers Using Standardized Test Scores. *European Journal of Education and Pedagogy*, 5(2), 1–11. <https://doi.org/10.24018/ejedu.2024.5.2.717>
12. Bernhofer, J., & Tonin, M. (2022). The effect of the language of instruction on academic performance. *Labour Economics*, 78, 102218. <https://doi.org/10.1016/j.labeco.2022.102218>
13. Aizawa, I. (2024). The role of language on assessment outcomes: an analysis of calculation and explanation questions in science classrooms. *Assessment & Evaluation in Higher Education*, 50(1), 67–82. <https://doi.org/10.1080/02602938.2024.2353865>
14. Ruegg, R., Hoang, H., & Petersen, N. (2024). The influence of English language proficiency test scores on the academic success of ESL undergraduate students. *Educational Research and Evaluation*, 29(3–4), 111–129. <https://doi.org/10.1080/13803611.2024.2314533>
15. Li, J. (2021). Perceived effects of CET4 test preparation, language ability, and test performance: An exploratory study of Chinese EFL learners. *Language Education & Assessment*, 4(2), 38–58. <https://doi.org/10.29140/lea.v4n2.480>
16. Cormier, D. C., Bulut, O., McGrew, K. S., & Kennedy, K. (2022). Linguistic Influences on Cognitive Test Performance: Examinee Characteristics Are More Important than Test Characteristics. *Journal of Intelligence*, 10(1), 8. <https://doi.org/10.3390/jintelligence10010008>
17. Barkaoui, K. (2025). The relationship between English language proficiency test scores and academic achievement: A longitudinal study of two tests. *Language Testing*, 42(3), 253–282. <https://doi.org/10.1177/02655322251319284>
18. Azkiyah, S. N., Rahmawati, Y., Shohibussirri, M., Nisa, Y. F., Sunengsih, N., & Mukminin, A. (2023). English proficiency as a predictor of academic performance in the context of non English as a Medium of Instruction. *PASAA*, 65, 32–56.
19. Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, 28, 24–34. <https://doi.org/10.1016/j.learninstruc.2013.04.001>
20. Henry, D. L., Nistor, N., & Baltes, B. (2014). Examining the relationship between math scores and English language proficiency. *Journal of Educational Research and Practice*, 4(1), 11–29. <https://doi.org/10.5590/JERAP.2014.04.1.02>
21. Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A metaanalysis. *Psychological bulletin*, 146(7), 595–634. <https://doi.org/10.1037/bul0000231>

22. Espinas, D. R., & Fuchs, L. S. (2022). The Effects of Language Instruction on Math Development. *Child development perspectives*, 16(2), 69–75.
<https://doi.org/10.1111/cdep.12444>
23. Nucette A, Hamamura T, Leitaos S, Biedermann B. Can learning a new language make you better at maths? A meta-analysis of foreign language learning and numeracy skills during early adolescence. *Bilingualism: Language and Cognition*. 2025;28(4):923-953. doi:10.1017/S1366728924000701
24. Vukovic, R. K., & Lesaux, N. K. (2013). The language of mathematics: investigating the ways language counts for children's mathematical development. *Journal of experimental child psychology*, 115(2), 227–244.
<https://doi.org/10.1016/j.jecp.2013.02.002>
25. S, M., & Mohammed, L. A. (2024). Mathematics Through a Linguistic Lens: The impact of English Language Proficiency in Solving Mathematical Word Problems Among Secondary ESL Students. *International Journal of Emerging Issues in Social Science, Arts, and Humanities*, 2(2), 46–57.
26. S, M., & Ali Mohammed, L. (2024). Mathematics Through a Linguistic Lens: The impact of English Language Proficiency in Solving Mathematical Word Problems Among Secondary ESL Students. *International Journal of Emerging Issues in Social Science, Arts and Humanities (IJEISSAH)*, 2(2), 46-57.
<https://doi.org/10.60072/ijeissah.2024.v2i02.005>
27. Chow, J. C., & Ekholm, E. (2019). Language domains differentially predict mathematics performance in young children. *Early Childhood Research Quarterly*, 46, 179–186. <https://doi.org/10.1016/j.ecresq.2018.02.011>
28. Kleemans, T., & Segers, E. (2020). Linguistic precursors of advanced math growth in first-language and second-language learners. *Research in developmental disabilities*, 103, 103661. <https://doi.org/10.1016/j.ridd.2020.103661>
29. Linnavalli, T., Jylänki, P., Kainulainen, J., Tervaniemi, M., & Törmänen, M. (2024). The associations between mathematical skills, cognitive performance, and language background in elementary school children. a two-year follow-up study. *Scandinavian Journal of Educational Research*, 1–15.
<https://doi.org/10.1080/00313831.2024.2419063>
30. Jourdain, L., & Sharma, S. (2016). Language challenges in mathematics education: A literature review. *Waikato Journal of Education*, 21(2).
31. González-Martín, A. M., Berd-Gómez, R., Saura-Montesinos, V., Biel-Maeso, M., & Abrahamse, E. (2024). On the Relationship Between Bilingualism and Mathematical Performance: A Systematic Review. *Education Sciences*, 14(11), 1172.
<https://doi.org/10.3390/educsci14111172>
32. Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics*, 36(5), 648–662.
<https://doi.org/10.1016/j.clinthera.2014.04.006>