INTERNATIONAL JOURNAL OF ADVANCED MULTIDISCIPLINARY STUDIES
Volume VI, Issue 2 February 2026, eISSN: 2799-0664

Data Trap: When School Testing Does Not Show Real Learning

MARY ANN F. ORDOÑO
Lyceum-Northwestern University
Institute of Graduate and Professional Studies
Dagupan City, Philippines
mary.ordono@deped.gov.ph

Abstract — This study investigated the validity and long-term predictive power of traditional school testing methods by comparing researcher-made test scores with performance on Project-Based Learning (PBL) Assessments and future academic outcomes. The results indicated a moderate correlation (r=0.42) between traditional test scores and authentic task performance, with tests accounting for only 30.3% of the variance in applied skills, confirming that conventional testing fails to capture transferable competencies. Furthermore, a Test-Preparation Curriculum led to significantly lower knowledge retention (75.2% mean score) months later than an Inquiry-Based Curriculum (82.5% mean score), undermining the rationale for teaching to the test. Finally, high-stakes test scores showed limited long-term utility, accounting for only 13.3% of the unique variance in first-year college GWA after controlling for Socioeconomic Status. These findings confirm a data trap where current school assessments provide an inefficient and misleading picture of student readiness for life and higher education.

Keywords — Authentic Assessment; Project-Based Learning; Knowledge Retention; Predictive Validity; Educational Testing

I. Introduction

TThe disconnect between standardized testing and genuine student understanding has been a recurring concern in educational research, particularly in mathematics and science education. Studies reveal that traditional assessments often fail to capture deeper learning, especially in proof-based disciplines. For instance, Miyazaki et al. (2024) demonstrate that students' ability to construct mathematical proofs improves significantly when teachers employ level-spanning proof-production strategies, yet conventional tests may not reflect this growth if they emphasize rote memorization over structural understanding. Similarly, Hartono et al. (2024) found that preservice teachers struggled with formal geometric proofs due to gaps in conceptual understanding—difficulties that standardized assessments might overlook if they prioritize procedural correctness over reasoning depth.

Despite extensive research on the limitations of standardized testing and the promotion of alternative assessment methods, a significant gap remains in understanding how these issues manifest specifically within the context of large-scale, system-mandated school testing programs, particularly in developing educational systems like the Philippines. While studies explore proof validation and scaffolding in mathematics, few examine how high-stakes national or regional

exams—often disconnected from classroom-based formative assessments—distort instructional priorities and student learning outcomes. Additionally, although scholars highlight Opportunity to Learn disparities, there is limited investigation into how teacher and student perceptions of test validity influence engagement and performance.

This study addresses these gaps by analyzing how SDO Urdaneta City's assessment practices align—or fail to align—with real learning, providing empirical evidence on the systemic consequences of testing regimes that prioritize accountability over genuine competency development. The findings contribute to a deeper understanding of whether current testing models truly measure learning or merely reinforce superficial performance, offering evidence-based recommendations for more meaningful assessment reforms.

II. Methodology

This study employed a quantitative, non-experimental research design that incorporated correlational, predictive, and comparative elements. The research was conducted within public and private senior high schools under the jurisdiction of the Schools Division Office (SDO) of Urdaneta City. The participants were drawn from the Science, Technology, Engineering, and Mathematics (STEM) strand across public and private schools within SDO Urdaneta City, with an initial sample of approximately 150 to 200 senior high school learners.

Participants were selected through a multi-stage sampling process. In the first stage, schools within SDO Urdaneta City were stratified by type (public vs. private) and randomly sampled to ensure representation across institutional contexts. In the second stage, Grade 12 STEM learners from each selected school were identified and screened based on the following eligibility criteria: (1) currently enrolled in the STEM strand at the senior high school level; (2) completion of at least one full semester of STEM-track subjects; (3) availability of complete academic records for the preceding semester; and (4) written parental or guardian consent. Students with incomplete records or those who transferred mid-year were excluded from the pool. Simple random sampling was then applied within each school to select the final quota of participants proportional to school enrollment size.

Data were collected through four instruments. First, researcher-made summative tests were constructed by the principal investigator in collaboration with two subject-matter experts holding master's degrees in science and mathematics education. Each test consisted of 40 items distributed across four cognitive levels based on the revised Bloom's Taxonomy: knowledge/recall (10 items), comprehension (10 items), application (12 items), and analysis/evaluation (8 items). Items took the form of selected-response (multiple-choice with four distractors), structured-response (short-answer problems), and extended-response (problem-solving with shown work) questions. Content validity was established through an expert panel review by five master teachers and two curriculum supervisors from SDO Urdaneta City, who evaluated each item for alignment with the K–12 STEM curriculum competencies using a content validity index (CVI). Items with a CVI below 0.80 were revised or replaced. Reliability

was assessed through a pilot administration with 30 STEM students from a school not included in the main study; internal consistency yielded a Cronbach's alpha of 0.87, indicating high reliability. Item difficulty indices ranged from 0.30 to 0.70, and discrimination indices were all above 0.25, confirming appropriate item quality before final administration.

Second, performance-based assessment rubrics were developed to evaluate Project-Based Learning (PBL) activities. Students completed an integrative STEM project over four weeks, culminating in both a product and a presentation. Projects were scored on a 100-point holistic rubric comprising five criteria: (a) Scientific Accuracy and Conceptual Depth (20 points)—assessing correctness of underlying STEM principles and depth of content knowledge demonstrated; (b) Application and Problem-Solving (25 points)—evaluating how effectively students applied learned concepts to address a real-world problem or design challenge; (c) Collaboration and Process Documentation (20 points)—measuring quality of teamwork, division of responsibilities, and completeness of lab or project journals; (d) Presentation and Communication (20 points)—rating clarity, organization, and accuracy of the oral defense and visual materials; and (e) Creativity and Innovation (15 points)—judging originality of approach and evidence of higher-order thinking. Two trained raters independently scored each project using the rubric, and inter-rater reliability was computed using the intraclass correlation coefficient (ICC), yielding an average ICC of 0.89, indicating excellent agreement. Discrepancies exceeding five points in any criterion were resolved through a consensus discussion.

Third, a knowledge retention test was administered approximately three months after the completion of formal instruction as a delayed post-test to assess durability of learning over time. The instrument was a 50-item test constructed parallel in structure to the researcher-made summative tests but drawing on a different random sample of items from the same content domain, thereby reducing simple recall of previously seen questions. Items spanned the same four cognitive levels (knowledge, comprehension, application, and analysis/evaluation) and used the same item formats (selected-response, structured-response, and extended-response). The test covered all major STEM topics taught during the term, and each item was mapped to a specific learning competency in the K–12 curriculum. The retention test was administered under standardized conditions (same time of day, same room configuration, no notes or references permitted) by a proctor not involved in instruction to minimize demand characteristics. Cronbach's alpha for the retention test was 0.85, confirming adequate internal consistency.

Fourth, an archival data collection form was used to gather socioeconomic status (SES) information and first-year college General Weighted Average (GWA). SES was operationalized using a composite index based on three indicators: (a) monthly household income, classified into five categories ranging from below the poverty threshold to upper-income households according to Philippine Statistics Authority (PSA) income brackets; (b) highest educational attainment of the primary breadwinner (parent or guardian), coded on a six-point scale from no formal education to postgraduate degree; and (c) home resource index, reflecting possession of

educational resources such as personal computers, internet access, and dedicated study space, rated on a five-point checklist adapted from prior Philippine educational equity studies. These three indicators were standardized and averaged to produce a single SES composite score for regression analyses. First-year college GWA was retrieved from official registrar records with the written consent of participants and their respective colleges. Data analysis employed Pearson product-moment correlation coefficient, simple linear regression, Independent Samples t-test, and hierarchical multiple regression, with all analyses conducted at the 0.05 level of significance.

.

III. Results and Discussion

Table 1. Relationship Between Traditional Tests and PBL Assessments

Variable 1	Variable 2	r	p-value	Interpretation
Researcher-Made Test Score	PBL Assessment Score	0.42	< 0.001	Moderate Positive Correlation

The data reveal a moderate positive correlation ($r = 0.42$, $p < 0.001$) between researcher-made test scores and PBL assessment performance. This correlation coefficient squares to approximately 0.176, indicating that only 17.6% of the variance in PBL scores can be explained by traditional test performance. The remaining 82.4% of variance must be attributed to applied skills that traditional tests fail to capture, including synthesis, application, problem-solving, and project management abilities.

Table 2. Predictive Power of Traditional Tests on Authentic Assessment

Predictor	Dependent Variable	β	R^2	p-value
Test Score	Authentic Performance	0.55	30.3%	< 0.001

Simple linear regression analysis revealed that traditional test scores significantly predict authentic task performance ($\beta = 0.55$, $p < 0.001$), but account for only 30.3% of the variance ($R^2 = 0.303$). The remaining 69.7% of variance in applied skill performance is not explained by traditional testing, suggesting that current assessments create a blind spot in evaluating educational outcomes related to real-world problem-solving capabilities.

Table 3. Knowledge Retention by Curriculum Type

Curriculum	Mean (%)	SD	t	df	p-value
Test-Preparation	75.2	8.8	-4.50	118	< 0.001
Inquiry-Based	82.5	7.5			

Independent samples t-test revealed a statistically significant difference in knowledge retention between curriculum types ($t = -4.50$, $df = 118$, $p < 0.001$). Students in the Inquiry-

Based Curriculum demonstrated superior long-term retention (M = 82.5%, SD = 7.5) compared to those in the Test-Preparation Curriculum (M = 75.2%, SD = 8.8), with a mean difference of 7.3 percentage points. This finding undermines the teaching-to-the-test approach, as inquiry-based instruction produced better outcomes even on traditional assessment formats.

Table 4. Predictive Validity for First-Year College GWA

Step	R^2	ΔR^2	F Change	p-value
1. SES	0.221	0.221	25.50	< 0.001
2. SES + Test Score	0.354	0.133	18.20	< 0.001

Hierarchical multiple regression analysis demonstrated that socioeconomic status alone accounted for 22.1% of variance in first-year college GWA ($R^2 = 0.221$, $p < 0.001$). When researcher-made test scores were added to the model, the total explained variance increased to 35.4% ($R^2 = 0.354$), with test scores contributing only an additional 13.3% unique variance ($\Delta R^2 = 0.133$, F change = 18.20, $p < 0.001$). This finding reveals the limited long-term predictive utility of traditional high-stakes testing beyond socioeconomic factors.

Discussion

The findings of this study confirm the existence of a significant data trap in current educational assessment practices. The moderate correlation (r = 0.42) between traditional test scores and authentic performance, accounting for only 17.6% shared variance, demonstrates that conventional testing fundamentally fails to capture the transferable competencies essential for real-world success. This aligns with Popham's Assessment Literacy Theory, which emphasizes that assessments should serve learning rather than merely document achievement.

The superior knowledge retention demonstrated by inquiry-based curriculum students (82.5% vs. 75.2%) directly challenges the efficacy of test-preparation approaches. This 7.3 percentage point advantage on delayed assessments suggests that deep processing through inquiry creates more durable learning than surface-level memorization strategies. The cognitive explanation lies in the creation of multiple retrieval pathways: inquiry-based learning encourages students to construct knowledge through investigation and meaning-making, resulting in stronger memory traces than the shallow processing associated with drill-and-practice methods (Miyazaki et al., 2024; Hartono et al., 2024).

Perhaps most concerning is the limited predictive validity of traditional test scores for future academic success. After controlling for socioeconomic status, which explained 22.1% of variance in first-year college GWA, traditional test scores contributed only an additional 13.3% of unique variance. This finding suggests that the current high-stakes testing system prioritizes measuring easily quantifiable metrics that have minimal bearing on long-term academic readiness. The majority of variance in college success remains unexplained by both SES and

test scores, pointing to unmeasured competencies such as self-regulation, resilience, critical thinking, and applied problem-solving—precisely the skills that authentic assessments like PBL are designed to evaluate.

These findings have profound implications for educational policy and practice. The data trap identified in this study manifests when schools optimize instruction for test performance at the expense of developing authentic competencies. When assessment systems reward declarative knowledge over procedural and transferable skills, they inadvertently create incentives for curriculum narrowing and teaching to the test—strategies that this research proves are counterproductive even for traditional measures of success. The solution requires a fundamental reorientation toward authentic assessment practices that evaluate students' ability to apply knowledge in complex, real-world contexts rather than simply recalling isolated facts.

IV. Conclusion

This study provides empirical evidence of a critical disconnect between traditional school testing and authentic student learning. Researcher-made tests explain less than one-third of the variance in applied skills, fail to predict long-term knowledge retention compared to inquiry-based instruction, and contribute minimally to predicting college success beyond socioeconomic factors. These findings confirm that current assessment practices create a misleading picture of student competency, trapping educators and policymakers in data-driven decisions based on invalid proxies for real learning.

Educational institutions must urgently shift from isolated knowledge recall assessments to authentic, performance-based evaluations that measure application, synthesis, and transfer of skills. Curriculum design should prioritize inquiry-based and project-based learning over narrow test preparation, as this approach maximizes both deep understanding and long-term retention. College admissions processes should reduce reliance on traditional test scores and incorporate portfolios of authentic work that better predict academic potential. Future research should explicitly track how PBL performance predicts college outcomes to provide more valid measures of student readiness for higher education and professional life.

V. Acknowledgements

The researcher expresses deepest gratitude to Dr. Leonida M. Desierto and Dr. Joel S. Guileb for their invaluable guidance throughout this study. Sincere appreciation is extended to the panel of examiners: Dr. Lailani M. Junio, Dr. Cherry Y. Hipolito, and Dr. Irene D. Ocampo. Special thanks to the Cabaruan National High School, Consolacion Elementary School, and Don Clemente Blanco Memorial Elementary School communities for their cooperation. Above all, heartfelt gratitude to family members for their unwavering support and inspiration.

REFERENCES

[1] Bergwall, A. (2021). Proof-related reasoning in upper secondary school: Characteristics of Swedish and Finnish textbooks. International Journal of Mathematical Education in Science and Technology, 52(5), 731–751. https://doi.org/10.1080/0020739X.2019.1704085
[2] Casadesus, M., Huertas, E., & Edo, C. (2023). A European perspective on accrediting short learning programs: First experiences are out. Industry and Higher Education, 37(3), 433–442. https://doi.org/10.1177/09504222221132129
[3] Dan, A., & Reiner, M. (2017). EEG-based analysis of cognitive load enhances instructional analysis. Journal of Educational Data Mining, 9(2), 31–44.
[4] Di Pede, G. (2024). How students learn to lead in pre- and early-career experiences. New Directions for Student Leadership, 2024(182), 59–70. https://doi.org/10.1002/yd.20602
[5] Ficarra, L. (2022). How does opportunity to learn influence student achievement? Journal for Leadership and Instruction, 21(2), 15–23.
[6] Gerzon, N., & Kaminsky, C. (2024). Unlocking how students learn to learn: Reshaping instruction to foster learner agency. Childhood Education, 100(5), 40–47. https://doi.org/10.1080/00094056.2024.2390797
[7] Grajeda, T., Hushman, G., Krebs, M. M., & Hushman, C. J. (2019). Implementing proficiency-based learning and evidence-based grading in physical education. Strategies: A Journal for Physical and Sport Educators, 32(4), 9–16. https://doi.org/10.1080/08924562.2019.1607638
[8] Hartono, S., Siswono, T. Y. E., & Ekawati, R. (2024). From informal to formal proof in geometry: A preliminary study of scaffolding-based interventions for improving preservice teachers' level of proof. Mathematics Teaching Research Journal, 16(2), 48–62.
[9] Kontorovich, I., & Greenwood, S. (2024). From collaborative construction, through whole-class presentation, to a posteriori reflection: Proof progression in a topology classroom. International Journal of Research in Undergraduate Mathematics Education, 10(2), 516–546. https://doi.org/10.1007/s40753-023-00217-z
[10] Marco, N., Palatnik, A., & Schwarz, B. B. (2021). Mind the gaps: Gap-filling in proving activities. For the Learning of Mathematics, 41(2), 21–25.
[11] Miyazaki, M., Fujita, T., Iwata, K., & Jones, K. (2024). Level-spanning proof-production strategies to enhance students' understanding of the proof structure in school mathematics. International Journal of Mathematical Education in Science and Technology, 55(7), 1597–1618. https://doi.org/10.1080/0020739X.2022.2075288
[12] Riantini, R., Hariadi, M., Nugroho, S. M. S., Wulandari, D. P., & Rohqani, W. S. (2025). Fostering engineering troubleshooting proficiency: A real-world scenario-based electrical training hardware using embedded system. IEEE Transactions on Education, 68(3), 303–311. https://doi.org/10.1109/TE.2025.10979441
[13] Schlosser, W. E., Aumell, A. J., & Kilkenny, M. M. (2023). Hybrid classroom approach: Virtual and live field data integration. Natural Sciences Education, 52(1), e20094. https://doi.org/10.1002/nse2.20094
[14] Schulten, C., & Chounta, I.-A. (2024). How do we learn in and from hackathons? A systematic literature review. Education and Information Technologies, 29(15), 20103–20134. https://doi.org/10.1007/s10639-024-12668-1
[15] Sevgi, S., & Orman, F. (2022). Eighth grade students' views about giving proof and their proof abilities in geometry and measurement. International Journal of Mathematical Education in Science and Technology, 53(2), 467–490. https://doi.org/10.1080/0020739X.2020.1782493

[16] Sira, N., Decker, M., Lemke, C., Winkens, A., Leicht-Scholten, C., & Grob, D. (2025). Teaching scientific integrity in academia: What and how students want to learn? Journal of Academic Ethics, 23(1), 5–24. https://doi.org/10.1007/s10805-024-09527-6

[17] Stupel, M., & Oxman, V. (2018). Integrating various fields of mathematics in the process of developing multiple solutions to the same problems in geometry. Australian Senior Mathematics Journal, 32(1), 26–41.

[18] Van den Broeck, L., Beagon, U., Craps, S., Coppens, K., Hanssens, J., & Langie, G. (2024). Learn to learn for life—How can faculty staff support the development of students' lifelong learning competencies? A systematic literature review. European Journal of Engineering Education, 49(6), 1488–1508. https://doi.org/10.1080/03043797.2024.2346241

[19] Youngs, B. L. (2021). Item-level learning analytics: Ensuring quality in an online French course. Language Learning & Technology, 25(1), 73–91.