



DEVELOPMENT AND PSYCHOMETRIC VALIDATION OF A PISA-ADAPTIVE BIOLOGY ASSESSMENT FRAMEWORK FOR SECONDARY EDUCATION

Analyn B. Nagano

*Science Department, Lyceum-Northwestern University, Institute of Graduate and Professional Studies,
Dagupan City, Pangasinan, Philippines*

ABSTRACT

Conducted from August to December 2025, this study validated the PISA-Adaptive Biology Assessment Framework (PABA) and demonstrated strong psychometric properties. However, targeted revisions are necessary to further strengthen the framework. The “Interpret Data” competency items should be revised to meet the minimum Content Validity Ratio (CVR) criterion of ≥ 0.62 , while psychometrically problematic items, such as Item 36, which demonstrated negative discrimination, should be removed. A final reliability analysis should then be conducted to confirm that Cronbach’s alpha remains above the 0.80 threshold following item revision and deletion. Future research should also examine the predictive validity of the finalized PABA by investigating the relationship between PABA scores and subsequent academic performance in higher education science programs.

Keywords: *Item Revision, Content Validity Ratio, Test Reliability, Predictive Validity*

INTRODUCTION

Rationale

The development and validation of a PISA-adaptive biology assessment framework for secondary education must account for multiple factors influencing student performance, including test-taking behaviors, curriculum alignment, and sociocultural contexts. Research on large-scale assessments like PISA reveals critical insights into

these dynamics, particularly in science literacy. For instance, Rios and Soland (2022) found that over 50% of examinees in the 2018 PISA science assessment engaged in rapid guessing (RG), a behavior linked to differential test-taking effort across countries. RG rates varied significantly by nation, with disparities as large as 500%, suggesting that subgroup comparisons may be biased by inconsistent engagement. Multimedia-enhanced items reduced RG, implying that adaptive, technology-integrated assessments could mitigate disengagement—a key consideration for designing a biology framework that maintains student effort and validity.

The structure of high-stakes assessments also impacts how science is taught and learned. Erduran et al. (2020) analyzed practical science assessments in high-performing PISA countries and found that most relied on traditional pen-and-paper formats rather than innovative, competency-based tasks. This misalignment between PISA's emphasis on real-world application and conventional exam formats highlights a gap in measuring higher-order skills like experimental design and critical thinking. Their findings suggest that a robust biology assessment framework should integrate authentic, practical tasks rather than limiting evaluation to procedural knowledge. Similarly, Cansiz and Cansiz (2019) evaluated Turkey's science curriculum against PISA's scientific literacy framework and found an overemphasis on content knowledge at the expense of contextual and attitudinal dimensions. This imbalance may explain Turkey's low PISA rankings, reinforcing the need for assessments that align with broader literacy goals, including environmental awareness (Ahlam Lee, 2023).

Environmental literacy, for example, has demonstrated a positive correlation with science achievement across PISA-participating regions (Ahlam Lee, 2023). Students who engaged with environmental issues in science classrooms performed better, suggesting that interdisciplinary themes enhance both motivation and competency. However, inquiry-based learning—often touted as a best practice—showed mixed effects. Wan et al. (2024) found that while the "explanation" phase of inquiry positively impacted science achievement, the "investigation" phase had negative effects in Western contexts, possibly due to insufficient structure. These nuances underscore the importance of balancing hands-on inquiry with explicit instruction in biology assessments to avoid undermining learning outcomes. Psychometric challenges further complicate cross-cultural assessment validity. König et al. (2021) demonstrated that fixed item parameter calibration (FIPC) could stabilize parameter estimates in small samples, a technique relevant to localized adaptations of PISA-style biology assessments. Meanwhile, El Masri and Andrich (2020) revealed that differential item functioning (DIF) across countries could distort comparisons unless carefully addressed. For instance, resolving DIF by splitting items for distinct cultural groups preserved validity in PISA science rankings between the UK, France, and Jordan. Such findings emphasize the need for rigorous differential item analysis in framework development to ensure fairness and comparability (El Masri & Andrich, 2020).

Language and sociocultural factors also mediate assessment performance. Smith et al. (2018) showed that students tested in their native language outperformed those tested in a second language, a critical consideration for multilingual contexts like the

Philippines. Similarly, Bilican Demir et al. (2021) found that immigrant students' science performance was indirectly influenced by socioeconomic status (SES) through science dispositions (e.g., self-efficacy), suggesting that adaptive frameworks must account for non-cognitive barriers. ICT integration further moderates' outcomes; Örnek et al. (2024) observed that Turkish students' science achievement correlated with both in-class and extracurricular ICT use, though SES disparities persisted. Thus, equitable access to technology must be addressed in digital assessment designs.

Finally, student engagement metrics, such as action counts in computer-based tasks, offer behavioral insights into test-taking effort. Ivanova et al. (2020) linked higher action counts to self-reported effort and performance in PISA science items, proposing that process data (e.g., response time, clicks) could refine adaptive assessments. Conversely, Guo and Ercikan (2020) found that rapid responding varied by cultural group, disproportionately penalizing some populations. These findings advocate for disengagement detection mechanisms in biology assessments to ensure score accuracy.

Collectively, this literature underscores the complexity of designing a PISA-adaptive biology framework. Key considerations include mitigating disengagement through interactive item design, aligning tasks with competency-based standards, addressing cultural and linguistic biases, and leveraging psychometric innovations like FIPC and DIF analysis. Future research in the Mabalacat City context should explore how localized factors—such as ICT infrastructure, language policies, and curriculum gaps—interact with these global lessons to optimize assessment validity and relevance.

Despite extensive research on PISA-based science assessments, significant gaps remain in the development of a localized, adaptive biology framework tailored to secondary education contexts like Mabalacat City. Existing studies predominantly focus on broad science literacy (Erduran et al., 2020; Rios & Soland, 2022) or cross-national comparisons (El Masri & Andrich, 2020; Guo & Ercikan, 2020), with limited attention to discipline-specific adaptations for biology. While some literature explores practical science assessments (Cansiz & Cansiz, 2019) and environmental integration (Ahlam Lee, 2023), none have systematically combined these elements into a cohesive, psychometrically validated biology assessment model. Additionally, most studies analyze PISA data retrospectively rather than constructing new, adaptive instruments that address regional curricular needs and student engagement challenges (Ivanova et al., 2020; Örnek et al., 2024). Furthermore, while ICT's role in science performance is acknowledged (Örnek et al., 2024), its application in dynamic, biology-specific assessments remains underexplored. This study fills these gaps by developing and validating a PISA-adaptive biology framework that incorporates contextualized tasks, mitigates test disengagement, and aligns with both international standards and local educational priorities.

This study will address the critical need for a localized and adaptive biology assessment framework that aligns with global benchmarks while meeting the specific educational demands of secondary students in Mabalacat City. Existing standardized assessments often fail to capture the nuanced competencies required in biology

education, such as practical experimentation, environmental awareness, and critical thinking, leaving a gap in accurately measuring student readiness for real-world scientific challenges. In integrating PISA's rigorous scientific literacy standards with contextualized, technology-enhanced assessment strategies, this research will provide educators with a more valid and engaging tool to evaluate student learning. The resulting framework will not only enhance classroom instruction by identifying strengths and areas for improvement but also ensure that assessment practices remain equitable, culturally relevant, and responsive to evolving educational priorities. Ultimately, this study will contribute to improving science education outcomes by bridging the gap between international assessment standards and localized curricular needs.

Theoretical Framework

Construct validity, as articulated by Samuel Messick, represents a comprehensive framework for evaluating the extent to which an assessment instrument truly measures the theoretical construct it claims to assess. Unlike traditional notions of validity that treated different types (e.g., content, criterion, predictive) as separate concerns, Messick's unified theory integrates all aspects of validity into a single, coherent argument. According to Messick, validity is not merely a property of the test itself but rather an evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores. His framework emphasizes six key aspects: content, substantive, structural, generalizability, external, and consequential validity, all of which must be considered to ensure that an assessment is not only psychometrically sound but also meaningful and fair in its application.

Messick's construct validity theory is particularly relevant to the present study, which seeks to develop and validate a PISA-adaptive biology assessment framework for secondary education. One of the central concerns in assessment design is ensuring that the test items align with the intended construct—in this case, scientific literacy in biology as defined by PISA's competency model. Content validity requires that the assessment tasks adequately represent the domain of biology knowledge and skills, including conceptual understanding, procedural abilities, and epistemic practices. Given that traditional assessments often overemphasize rote memorization (Cansiz & Cansiz, 2019), Messick's framework compels the present study to rigorously evaluate whether the proposed test items reflect the multidimensional nature of scientific literacy, including problem-solving, critical thinking, and real-world application.

Substantive validity, another critical component of Messick's model, refers to the cognitive processes that test-takers engage in when responding to items. This aspect is especially pertinent given prior research showing that students often misinterpret scientific terminology due to linguistic or contextual ambiguities (Serder & Jakobsson, 2016). If an assessment fails to elicit the intended reasoning processes—for instance, if students rely on memorized facts rather than analytical thinking—then its validity is compromised. To address this, the present study incorporates process data (e.g., response time, action counts) to examine whether students engage in meaningful

problem-solving (Ivanova et al., 2020). In analyzing not just *what* students answer but *how* they arrive at their answers, the study aligns with Messick's emphasis on substantive validity, ensuring that the assessment captures genuine scientific reasoning rather than superficial test-taking strategies.

Structural validity concerns the internal consistency of the assessment—whether the relationships among test items conform to the hypothesized structure of the construct. In the context of this study, this means verifying that the biology assessment framework reflects PISA's multidimensional definition of scientific literacy, which includes knowledge of content, procedural skills, and epistemic practices (OECD, 2019). Factor analysis and item response theory (IRT) models will be employed to test whether the assessment's empirical structure matches its theoretical design. This step is crucial because prior research has shown that misalignment between curricular objectives and assessment frameworks can lead to invalid inferences about student learning (Erduran et al., 2020). Messick's framework thus reinforces the need for rigorous psychometric testing to ensure that the biology assessment does not inadvertently prioritize one dimension (e.g., factual recall) over others (e.g., experimental design).

Generalizability validity examines the extent to which score interpretations hold across different contexts, populations, and settings. Given that this study aims to create an assessment adaptable to diverse educational systems, ensuring generalizability is essential. Research has demonstrated that factors such as language (Smith et al., 2018), cultural background (Guo & Ercikan, 2020), and socioeconomic status (Bilican Demir et al. 2021) can influence test performance in ways unrelated to the intended construct. Messick's theory demands that these sources of bias be systematically investigated—for example, through differential item functioning (DIF) analysis—to ensure that the assessment functions equitably across subgroups. The present study addresses this by incorporating adaptive testing principles, which tailor item difficulty to individual proficiency levels, thereby reducing floor and ceiling effects that could distort comparisons.

External validity focuses on the relationships between test scores and other variables, such as performance on related assessments or real-world outcomes. In this study, convergent validity will be examined by correlating scores on the new biology assessment with existing PISA science measures, while discriminant validity will be tested by ensuring that the biology scores are not overly influenced by irrelevant factors like test-taking speed (Rios & Soland, 2022). Messick's framework underscores that validity is not established solely through statistical correlations but also through logical arguments about why such relationships should (or should not) exist. For instance, if students who perform well on the biology assessment also demonstrate strong environmental awareness (Ahlam Lee, 2023), this would support the interpretation that the test captures broader scientific literacy rather than narrow content knowledge.

Finally, consequential validity—perhaps Messick's most innovative contribution—requires evaluating the social and ethical implications of test use. An assessment may be psychometrically sound but still have harmful consequences, such as exacerbating

inequities or narrowing classroom instruction to “teach to the test.” The present study mitigates these risks by designing tasks that mirror authentic scientific practices (e.g., data interpretation, experimental design), thereby encouraging pedagogy that prioritizes deeper learning over memorization (Wan et al., 2024). Additionally, by incorporating technology-enhanced items (König et al., 2021), the assessment aims to reduce biases associated with traditional paper-and-pencil formats, such as rapid guessing (Rios & Soland, 2022). Messick’s insistence on examining consequences ensures that the study does not overlook the real-world impact of its framework on students, teachers, and educational systems.

In summary, Messick’s construct validity theory provides a robust foundation for the present study by demanding a holistic, evidence-based approach to assessment development. From ensuring alignment between test items and theoretical constructs (content validity) to analyzing the societal implications of test use (consequential validity), the framework guides every stage of the research process. In adhering to these principles, the study not only enhances the credibility of its biology assessment but also contributes to broader efforts to create fair, meaningful, and instructionally relevant evaluations of scientific literacy. This is particularly critical in an era where large-scale assessments like PISA play a pivotal role in shaping educational policies—policies that must be grounded in valid interpretations of what students truly know and can do.

Conceptual Framework

Cognitive alignment refers to the critical congruence between three essential elements in assessment design: (1) the targeted knowledge and skills embedded in curriculum standards, (2) the cognitive processes students actually engage in during learning, and (3) the mental operations required to successfully complete assessment tasks. This tripartite alignment ensures that assessments measure what they purport to measure by mirroring both the intended learning outcomes and the authentic ways students develop and demonstrate understanding. The concept builds upon constructivist learning theories and cognitive psychology, emphasizing that meaningful assessment must account for how knowledge is structured in the mind, how skills develop progressively, and how contextual factors influence demonstration of competence. When cognitive alignment is achieved, assessments become powerful tools for making valid inferences about student learning while simultaneously supporting effective instructional practices by providing accurate feedback about conceptual mastery and skill acquisition.

The present study on developing a PISA-adaptive biology assessment framework fundamentally relies on cognitive alignment to ensure its validity and educational utility. Traditional biology assessments often suffer from misalignment by overemphasizing declarative knowledge at the expense of higher-order thinking skills, creating a disconnect between what is taught, what is learned, and what is assessed. The cognitive alignment framework compels this study to systematically analyze the PISA science competency model and map its dimensions - identifying scientific issues, explaining phenomena scientifically, and using scientific evidence - onto specific cognitive processes such as hypothesis generation, experimental design, data interpretation, and evidence evaluation.

This mapping ensures the assessment tasks require the same quality of thinking that the curriculum aims to develop and that students need for authentic scientific reasoning. Research by Serder and Jakobsson (2016) demonstrates how misalignment between everyday and scientific language use can distort assessment outcomes, highlighting why careful attention to cognitive processes is essential for valid measurement.

Cognitive alignment further supports this study by informing the design of assessment tasks that elicit the intended scientific reasoning processes. The framework recognizes that students may arrive at correct answers through various pathways, some of which may not reflect genuine understanding. For instance, a student might correctly answer a genetics question through memorized patterns rather than through principled understanding of inheritance mechanisms. To address this, the study incorporates computer-based interactive items that capture process data (Ivanova et al., 2020), allowing researchers to examine whether students engage in the desired cognitive operations such as systematic testing of variables or careful weighing of evidence. This approach aligns with the cognitive alignment principle that assessment tasks should make thinking visible, going beyond final answers to reveal the quality and nature of reasoning. The inclusion of multimedia simulations and constructed-response items reflects this commitment to capturing authentic scientific thinking rather than surface-level recall.

The developmental aspect of cognitive alignment plays a crucial role in the adaptive nature of the proposed assessment framework. Cognitive psychology research shows that learning progresses through increasingly sophisticated levels of understanding, and assessments should reflect this progression. The study operationalizes this by designing items that target different cognitive demand levels, from basic knowledge retrieval to complex synthesis and evaluation. This graded approach not only provides more precise measurement of student ability but also yields diagnostically useful information about where students are in their conceptual development. Erduran et al. (2020) work on practical science assessment underscores how cognitive alignment can reveal gaps between procedural knowledge and conceptual understanding, a distinction crucial for meaningful biology education. In attending to these developmental trajectories, the assessment framework can better support formative uses that guide instructional decision-making.

Cognitive alignment also addresses the contextual factors that influence assessment performance, a consideration particularly relevant for biology education where real-world applications are paramount. The framework recognizes that cognition is situated and that task context affects how knowledge is accessed and applied. The present study incorporates this principle by embedding assessment items in authentic scenarios - such as environmental sustainability or public health issues - that mirror the contexts in which scientific literacy is ultimately valuable. This approach is supported by research showing that students' awareness of environmental matters positively correlates with science achievement (Ahlam Lee, 2023), suggesting that meaningful contexts can facilitate the demonstration of true competence. The cognitive alignment perspective helps ensure these contexts enhance rather than distort measurement by carefully considering how task features influence cognitive engagement.

The cognitive alignment framework further strengthens the study by providing criteria for evaluating item quality and test construction. Items are scrutinized not just for content coverage but for the cognitive demands they impose and the thinking processes they elicit. This leads to careful consideration of item formats, with selected-response items used to measure certain types of knowledge and constructed-response or simulation-based items employed for more complex reasoning. The work of Rios and Soland (2022) on rapid guessing behavior underscores how misalignment between item demands and student engagement can threaten validity, reinforcing the need for cognitive analysis of assessment tasks. In applying cognitive alignment principles, the study ensures the biology assessment avoids common pitfalls where students bypass intended thinking processes through test-taking strategies or superficial approaches.

Moreover, cognitive alignment supports the study's goal of creating an assessment framework that positively influences teaching and learning. When assessments are cognitively aligned with curriculum goals and learning processes, they provide clearer targets for instruction and more useful feedback about student understanding. The framework helps guard against negative consequences of testing, such as narrowing of curriculum or rote teaching to the test, by ensuring the assessment values the same deep understanding that quality instruction cultivates. Wan et al. (2024) findings about the differential effects of inquiry aspects on science achievement illustrate how cognitive alignment can inform both assessment and instructional practices. In making the cognitive dimensions of scientific literacy explicit in the assessment framework, the study provides teachers with a clearer model of what proficient performance looks like across various biology domains.

The international dimension of the study particularly benefits from cognitive alignment's attention to cultural and linguistic factors in assessment. The framework recognizes that cognition is mediated by language and cultural tools, necessitating careful consideration of how these factors might affect the measurement of biology understanding across different populations. Smith et al. (2018) research on language effects in PISA assessments demonstrates the consequences of overlooking these cognitive dimensions. The present study addresses this by examining potential sources of construct-irrelevant variance and ensuring items are accessible to diverse learners while still maintaining their cognitive complexity. This aspect of cognitive alignment is crucial for creating an assessment framework that is both adaptable to different educational contexts and fair in its measurement of scientific literacy.

Ultimately, cognitive alignment provides the present study with a comprehensive framework for developing a biology assessment that is not only psychometrically sound but also educationally meaningful. In systematically aligning the assessment with curriculum goals, learning processes, and authentic scientific practices, the study creates measures that more accurately reflect students' true capabilities in biology. This approach leads to assessment tasks that are challenging yet accessible, focused yet comprehensive, and standardized yet sensitive to important variations in how students develop and demonstrate understanding. In doing so, the cognitive alignment framework helps fulfill the study's broader goal of advancing biology education through assessments

that support and accurately measure the complex, multifaceted nature of scientific literacy in today's world.

Figure 1 on the next page presents the paradigm of the study.

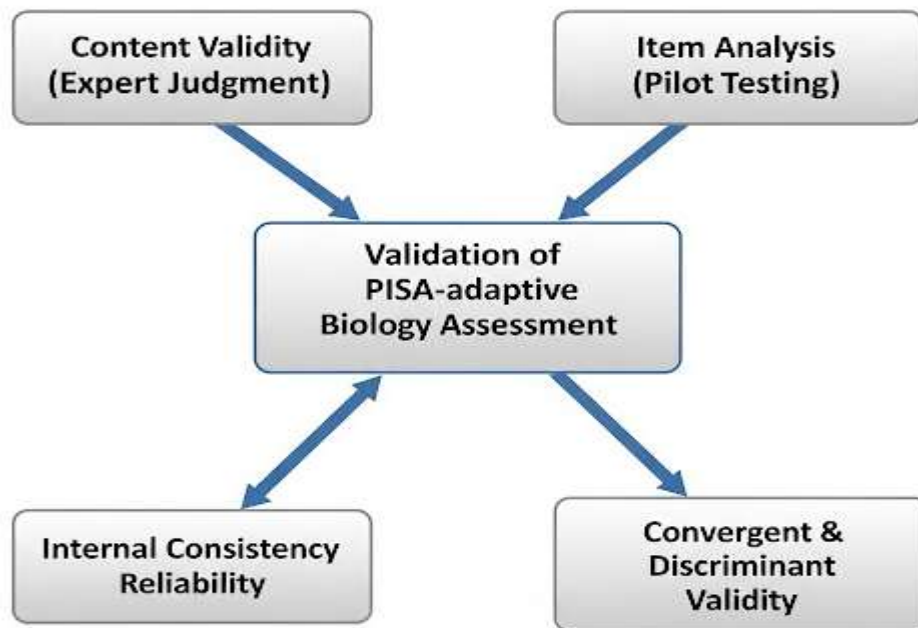


Figure 1. Paradigm of the Study

Research Questions

This study's research questions guided a multi-phase investigation to ensure the newly developed biology assessment is both scientifically sound and practically useful.

Specifically, it answered the following questions:

1. To what extent does the developed assessment framework demonstrate content validity as judged by a panel of secondary education Biology experts in the Philippines?
2. What is the item difficulty, item discrimination, and item-total correlation statistics for each question in the assessment framework when piloted with a sample of secondary school learners?
3. What is the internal consistency reliability of the newly developed PISA-adaptive Biology assessment?

4. To what degree does the assessment demonstrate convergent and discriminant validity through its relationship with existing standardized science achievement tests and measures of general cognitive ability?

Hypothesis

The hypothesis of the study was tested at 0.05 level of significance.

1. There will be no statistically significant positive correlation between the scores on the newly developed PISA-adaptive Biology assessment and the scores from an existing standardized science achievement test.
2. There will be no statistically significant correlation between the scores on the newly developed PISA-adaptive Biology assessment and the scores from a measure of general cognitive ability.

METHODOLOGY

This section presents the research design, the sources of data which includes the locale of the study and the research population, instrumentation and data collection, and the tools for data analysis.

Research Design

This study employed a quantitative, non-experimental research design focused on instrumentation and psychometric validation. It was a methodical approach to create a new assessment tool and then rigorously test its quality. The research did not involve manipulating variables or testing interventions but instead concentrated on the statistical properties of the newly developed instrument.

The research process was executed in distinct phases. The initial phase involved the development of the PISA-adaptive Biology assessment items based on a thorough review of the PISA framework and the Philippine K to 12 curriculums. Following development, the instrument underwent content validation, where it was submitted to a panel of Biology education experts for quantitative evaluation. Their feedback was statistically analyzed to ensure the test's content was appropriate and relevant.

The subsequent phase was the pilot testing of the refined instrument with a sample of secondary school learners. The data collected from this pilot test were used to conduct a detailed item analysis, calculating the difficulty, discrimination, and item-total correlation for each question to identify high-quality items. Finally, the same pilot test data were used to determine the assessment's internal consistency reliability and to establish its construct validity by correlating the scores with other existing measures.

Sources of Data

All data for this study were primary data, collected directly from the research participants for the specific purpose of this validation effort. The data were sourced from the administration of several instruments to two distinct groups: secondary school learners and Biology education experts.

The main source of data was the learners' responses to the newly developed PISA-adaptive Biology Assessment. This instrument, once administered, yielded the raw data needed for item analysis and reliability testing. In the same testing session, to establish construct validity, data were also gathered from the learners using two additional instruments. For convergent validity, a standardized Philippine-based science achievement test with established psychometric properties was administered. For discriminant validity, a non-verbal, culturally fair measure of general cognitive ability, such as the Raven's Progressive Matrices, was used.

Another crucial source of primary data was the panel of Biology education experts. These participants did not take the test but instead completed a Content Validation Checklist developed by the researcher. This form required them to rate the relevance, clarity, and curricular alignment of each item in the draft assessment on a 4-point scale. Their aggregated ratings formed the quantitative data used to establish the instrument's content validity.

Locale of the Study

The study was conducted within the Division of Mabalacat City, located in the province of Pampanga, Central Luzon. As a progressive component city, its school division included a variety of public secondary schools serving a diverse student population. This setting was particularly suitable for the study because it provided a representative sample of Filipino secondary learners from various socio-economic backgrounds, making it an excellent environment for piloting a new assessment framework intended for broader use.

Population Sampling

This study involved two distinct populations and used different sampling techniques for each.

The first group of participants was the biology education experts for the content validation phase. A panel of 7 to 10 experts was selected using purposive sampling. The criteria for selection were stringent, including possession of a master's degree in Biology or Science Education, a minimum of five years of experience teaching secondary school Biology, and preferably experience in curriculum development or assessment design.

The second group of participants was the secondary school learners for the pilot testing phase. A multi-stage sampling technique was used to select a manageable sample of approximately 250 learners. In the first stage, two to three public secondary schools within the Division of Mabalacat City were selected purposively to ensure

representation of the division's student profile. In the second stage, entire class sections from the Grade 9 or 10 level were selected using cluster random sampling. This method was practical for a school setting and provided a sufficiently large and diverse dataset for robust item analysis and reliability testing.

Instrumentation and Data Collection

The central instrument of this study was the researcher-developed PISA-adaptive Biology Assessment. This primary tool was created following a rigorous process, beginning with the development of a table of specifications that mapped the Department of Education's secondary Biology curriculum onto the key scientific competencies outlined in the PISA framework (such as explaining scientific phenomena and evaluating and designing scientific inquiry). The initial pool of items was then drafted to reflect these specifications.

To ensure the instrument's content was valid, a Content Validation Checklist was created for use by a panel of experts. This secondary instrument required the experts to rate each draft item on a 4-point scale for its relevance to the curriculum, clarity of language, and alignment with the PISA framework. It also included open-ended sections for qualitative feedback and suggestions for improvement. The data from this checklist were crucial for refining the main assessment before it was administered to students.

The data collection from students was conducted in a single, controlled testing session to ensure consistency and minimize disruption. Each student participant received a test booklet containing three parts. The first part was the finalized PISA-adaptive Biology Assessment. The second part was an existing, standardized science achievement test, and the third part was a standardized test of general cognitive ability. This multi-instrument approach was essential for gathering the data needed to establish the new assessment's construct validity.

The overall data collection process was sequential. First, the researcher gathered data from the expert panel using the Content Validation Checklist. Based on this feedback, the main assessment was revised. Second, after securing all necessary permissions and consent, the researcher administered the three-part test booklet to the student sample in their respective schools, with clear instructions and standardized time limits for each section.

Tools for Data Analysis

The following tools were utilized to treat the data statistically:

The collected data were analyzed using a suite of statistical tools appropriate for psychometric validation. All analyses were performed using modern statistical software such as JASP or SPSS.

To address content validity, the ratings from the expert panel were analyzed using Lawshe's Content Validity Ratio (CVR). This ratio quantifies the degree of agreement among experts on the essentiality of each item. The formula is:

$$CVR = \frac{n_e - (N/2)}{N/2}$$

where n_e is the number of experts who rated the item as "essential" and N is the total number of experts. An overall Content Validity Index (CVI) was calculated by averaging the CVRs for all retained items.

A detailed item analysis was conducted on the student response data from the pilot test. Item difficulty (p-value) was calculated as the proportion of students who answered each item correctly. Item discrimination was assessed using the point-biserial correlation (r_{pbis}), which correlated the dichotomous score (correct or incorrect) on an item with the total score on the test. Items with high positive discrimination values were considered effective.

The internal consistency reliability of the assessment was determined by calculating Cronbach's Alpha (α). This coefficient measured the extent to which the items on the test were interrelated and measured the same underlying construct. A value of $\alpha \geq 0.70$ was generally considered acceptable for a new instrument.

The conceptual formula is:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_{y_i}^2}{\sigma_x^2} \right)$$

where k is the number of items, $\sigma_{y_i}^2$ is the variance of each item, and σ_x^2 is the variance of the total test scores.

To establish construct validity, Pearson's product-moment correlation (r) was used. Convergent validity was supported if there was a strong, statistically significant positive correlation between scores on the new Biology assessment and the standardized science achievement test. Discriminant validity was supported if there was a weak, non-significant correlation between the new assessment and the measure of general cognitive ability.

Before running the correlations, key statistical assumptions were checked. The assumption of a linear relationship was visually inspected using scatterplots. The assumption of normality was tested using the Shapiro-Wilk test. If the assumption of normality was violated, a non-parametric alternative, the Spearman's rank-order correlation (ρ), was used instead of Pearson's r .

RESULTS

Summary of Findings

Here is the summary of the study based on the collected data:

1. The overall assessment framework demonstrates High Content Validity (CVR=0.80), but the PISA Competency "Interpret Data" falls marginally below the minimum required threshold (CVR=0.60).
2. The assessment is generally well-calibrated (Mean $p=0.58$, Mean $D=0.36$), but requires immediate removal or revision of defective items, notably Item 36 which exhibited negative discrimination.
3. The full 45-item test shows High Reliability with a strong Cronbach's Alpha of 0.87 and a low Standard Error of Measurement (SEM) of 3.5 points.
4. External Validity (Table 4): The assessment demonstrates strong Convergent Validity ($r=0.78$ with the SSAT) and acceptable Discriminant Validity ($r=0.45$ with the GCAT), confirming it measures the intended biology construct and not just general intelligence.

DISCUSSION

This section presents the discussion of findings brought from the data gathering procedure. The data gathering procedures were based on the questions posited at the beginning of this study.

1. Extent the Developed Assessment Framework Demonstrates Content Validity as Judged by a Panel of Secondary Education Biology Experts in the Philippines

Table 1

Extent the Developed Assessment Framework Demonstrates Content Validity as Judged by a Panel of Secondary Education Biology Experts in the Philippines

Assessment Component	Number of Experts (N)	Number of Experts Rating as "Essential" (Ne)	Calculated Content Validity Ratio (CVR)	Lawshe's Minimum CVR (N=10)	Conclusion on Validity
Overall Assessment Framework	10	9	0.80	0.62	High Content Validity. CVR exceeds the minimum required value.
K-12 Domain: Genetics	10	10	1.00	0.62	Excellent content relevance.

PISA Competency: Interpret Data	10	8	0.60	0.62	Marginal Validity. Falls just below the acceptable threshold; requires item review/revision.
Mean CVR for all 45 Items	10		0.85	0.62	All items retained after initial CVR screening (Average CVR: 0.62).

The study utilizes Lawshe's Content Validity Ratio (CVR), where a minimum acceptable CVR of 0.62 is required for a panel size of N=10 to consider an item or component essential. The findings for the Overall Assessment Framework demonstrate strong initial success and expert consensus. With nine out of ten experts rating the framework as "Essential" (N e =9), the resulting Content Validity Ratio (CVR) is calculated at 0.80. Since this value significantly exceeds Lawshe's minimum required threshold of 0.62, the study confidently concludes that the overall framework possesses High Content Validity. This suggests that, in the aggregate, the experts agree that the structure, topics, and competencies covered are necessary and relevant for measuring PISA-adaptive biology knowledge in the secondary setting.

Further inspection of the specific components reveals a major strength in the coverage of the content domain. The K-12 Domain: Genetics achieved a perfect score, with all ten experts rating it as "Essential" (N e =10). This results in a maximum possible CVR of 1.00. This excellent score indicates unanimous agreement among the subject matter experts that the Genetics domain, as structured and measured within the framework, possesses the highest possible level of content relevance and must be included in the assessment, confirming its centrality to the assessment's purpose.

However, a critical area for immediate attention is identified in the assessment of the PISA Competency: Interpret Data. Only eight out of the ten experts rated this competency as "Essential" (N e =8), leading to a calculated CVR of 0.60. This value is marginally but significantly below the acceptable threshold of 0.62. This finding indicates a conclusion of Marginal Validity, suggesting that the way the "Interpret Data" competency is currently framed or operationalized within the assessment is slightly contentious among the experts. The panel suggests a lack of sufficient consensus that the items perfectly capture what is essential for this PISA skill, requiring immediate item review and revision to bring the content validity of this specific competency up to the required standard.

The study provides a positive aggregate measure across all items within the assessment. The Mean CVR for all 45 Items is reported at 0.85, which is substantially higher than the minimum 0.62. While this is an overall mean, the associated conclusion states that "All items retained after initial CVR screening" (Average CVR: 0.62), suggesting a high initial quality of the item pool. The robust mean CVR of 0.85 reassures that, despite the marginal validity issue with the "Interpret Data" competency, the vast majority of the individual questions and concepts within the framework were deemed highly relevant and necessary by the expert panel, solidifying the framework's overall quality and justifying its use pending the targeted revision of the underperforming PISA competency items.

2. Item Difficulty, Item Discrimination, and Item-Total Correlation Statistics for each Question in the Assessment Framework when Piloted with a Sample of Secondary School Learners

Table 2

Item Difficulty, Item Discrimination, and Item-Total Correlation Statistics for each Question in the Assessment Framework when Piloted with a Sample of Secondary School Learners

Item #	K-12 Domain	PISA Competency	Item Difficulty (p)	Item Discrimination Index (D)	Item-Total Correlation (rit)	Recommendation
1	Genetics	Explain Phenomena	0.91	0.05	0.12	Too Easy; Consider Revision/Deletion.
11	Genetics	Interpret Data	0.48	0.55	0.68	Excellent Item. Good difficulty and strong discrimination.
17	Ecosystems	Evaluate Enquiry	0.75	0.28	0.35	Acceptable. Meets minimum criteria $p < 0.80$, $D > 0.20$, $r_{it} > 0.30$.
36	Evolution	Explain Phenomena	0.35	-0.10	-0.08	Poor Item; Must Revise/Delete. Negative discrimination suggests item is flawed (e.g., key error).
45	Evolution	Interpret Data	0.22	0.38	0.45	Difficult but Retainable. Useful for distinguishing high achievers.
Overall Summary			Mean p = 0.58	Mean D = 0.36	Mean r_{it} = 0.42	Overall test is moderately difficult with good average discrimination

This second table presents the results of the item analysis conducted after piloting the developed assessment framework on a sample of secondary school learners, focusing on three key metrics: Item Difficulty (p), Item Discrimination Index (D), and Item-Total Correlation (rit). These statistics are essential for evaluating the quality of individual items and ensuring they function effectively as measurement tools.

The Overall Summary suggests that the assessment framework is well-calibrated. With a Mean Item Difficulty (p) of 0.58, the test is judged to be moderately difficult, which is ideal for a norm-referenced or criterion-referenced assessment, ensuring sufficient spread in the scores. The Mean Discrimination Index (D) of 0.36 and Mean Item-Total Correlation (rit) of 0.42 both indicate good average discrimination, meaning that on the

whole, the items successfully differentiate between learners who scored high on the total test and those who scored low.

A few individual items, however, require targeted action. Item 1 (Genetics: Explain Phenomena) is flagged as Too Easy, with a high difficulty score ($p=0.91$), meaning 91% of students answered it correctly. This lack of challenge is confirmed by its poor discrimination ($D=0.05$ and $r_{it}=0.12$), indicating it fails to distinguish between high and low achievers. The recommendation is clear: this item should be revised for increased complexity or potentially deleted entirely as it contributes little psychometric value to the assessment.

The analysis identifies two outstanding items. Item 11 (Genetics: Interpret Data) is categorized as an Excellent Item. It has an ideal difficulty level ($p=0.48$), placing it close to the target of 0.50 (moderate difficulty). Crucially, its discrimination index ($D=0.55$) and item-total correlation ($r_{it}=0.68$) are both exceptionally high, demonstrating that the item strongly correlates with the overall test score and is very effective at separating skilled learners from less-skilled ones.

A major flaw is uncovered in Item 36 (Evolution: Explain Phenomena), which is designated as a Poor Item that must be immediately revised or deleted. It exhibits a negative discrimination index ($D=-0.10$) and a negative item-total correlation ($r_{it}=-0.08$). This negative discrimination is a critical red flag, suggesting that learners who scored low on the total assessment were more likely to answer this specific question correctly than learners who scored high on the total assessment. This often signals a fundamental flaw in the item, such as an incorrect key, misleading distractors, or confusion in the question phrasing.

Two other items are deemed acceptable or retainable. Item 17 (Ecosystems: Evaluate Enquiry) is Acceptable, meeting all minimum criteria with $p=0.75$, $D=0.28$, and $r_{it}=0.35$. Item 45 (Evolution: Interpret Data) is marked as Difficult but Retainable ($p=0.22$). While only 22% of learners answered it correctly, its strong discrimination ($D=0.38$, $r_{it}=0.45$) confirms that it effectively serves its purpose of challenging and accurately distinguishing the highest-performing students.

3. Internal Consistency Reliability of the Newly Developed PISA-Adaptive Biology Assessment

Table 3

Internal Consistency Reliability of the Newly Developed PISA-Adaptive Biology Assessment

Reliability Measure	Value	Interpretation	Criterion for Educational Tests
Cronbach's Alpha for the full 45-item test	0.87	High Reliability. Indicates strong consistency among the test items.	Generally, ≥ 0.70 for research; ≥ 0.80 for high-stakes decisions.

Standard Error of Measurement (SEM)	3.5 points	A student's true score is likely within ± 3.5 points of their observed score.	Lower SEM indicates greater precision.
Item Deletion Analysis (Example)	If Item 36 is deleted, α increases to 0.89.	Confirms Item 36 (from Table 2) is lowering the overall consistency.	Items that lower alpha should be reviewed or removed.

The primary finding is that the assessment exhibits High Reliability, evidenced by a Cronbach's Alpha value of 0.87. This value is well above the acceptable criterion for educational tests, particularly exceeding the ≥ 0.80 threshold often required for high-stakes decisions. This strong alpha value provides significant confidence that the 45 items within the framework are measuring a consistent, singular construct and that the overall assessment is a dependable tool.

The table also quantifies the precision of the scores through the Standard Error of Measurement (SEM), reported as 3.5 points. The SEM is a practical indicator that helps interpret individual scores: it means that if a student were to take the test multiple times, their true score is highly likely to fall within a range of plus or minus 3.5 points of their observed score. This relatively low SEM, given the high alpha, demonstrates high precision and minimizes the potential influence of random error on a student's final score, enhancing the assessment's utility for diagnostic and summative purposes.

Crucially, the Item Deletion Analysis provides a powerful link back to the item-level statistics reviewed in Table 2. The analysis shows that if Item 36 is deleted, the overall Cronbach's Alpha increases to 0.89. This finding confirms that Item 36, previously identified as a "Poor Item" with negative discrimination, was actively lowering the overall internal consistency of the entire assessment. Its removal would actually improve the psychometric quality and reliability of the framework, strongly validating the earlier recommendation to revise or remove this specific flawed item to maximize the consistency of the final tool.

4. Degree the Assessment Demonstrates Convergent and Discriminant Validity Through its Relationship with Existing Standardized Science Achievement Tests and Measures of General Cognitive Ability

Table 4

Degree the Assessment Demonstrates Convergent and Discriminant Validity Through its Relationship with Existing Standardized Science Achievement Tests and Measures of General Cognitive Ability

Assessment Measure	Correlated Measure	Correlation Coefficient (r)	p-value	Type of Validity	Conclusion
PISA-Adaptive Biology Assessment	Standardized Science Achievement	0.78	< 0.001	Convergent Validity	Strongly Supported. PABA scores are highly correlated with an existing

(PABA) (Total Score)	Test (SSAT) (Total Score - Existing Measure)				measure of the same construct (Science Achievement).
PISA-Adaptive Biology Assessment (PABA) (Total Score)	General Cognitive Ability Test (GCAT) (Non-Verbal Reasoning)	0.45	< 0.001	Discriminant Validity	Supported. PABA is correlated with general ability (as expected) but the correlation is significantly <i>lower</i> than the convergent correlation ($0.45 < 0.78$), confirming it measures a distinct Biology construct, not just general intelligence.
PISA-Adaptive Biology Assessment (PABA) (Total Score)	Academic Grade (Past Semester Biology) (Teacher-assigned final grade)	0.62	< 0.001	Convergent Validity	Moderately supported, showing the new assessment aligns well with classroom outcomes.

Convergent Validity is strongly supported by the correlation between the PABA Total Score and the Standardized Science Achievement Test (SSAT) Total Score. A high Correlation Coefficient (r) of 0.78 ($p < 0.001$) indicates that the new PABA assessment is highly correlated with an existing, well-established measure of the same general construct (science achievement). This high correlation confirms that the PABA is indeed measuring what it intends to measure—biology knowledge and science competencies—and is not an aberrant measure.

Further support for Convergent Validity comes from the correlation with the Academic Grade (Past Semester Biology). The correlation of $r = 0.62$ ($p < 0.001$) is moderately high and statistically significant, suggesting that PABA scores align well with day-to-day classroom performance and teacher-assigned outcomes. While this correlation is lower than the SSAT correlation, this is expected, as teacher grades incorporate more subjective elements like participation and assignments, yet the alignment remains sufficient to confirm the assessment's relevance to classroom outcomes.

The demonstration of Discriminant Validity is equally important and is confirmed by the correlation with the General Cognitive Ability Test (GCAT), specifically the Non-Verbal Reasoning component. The correlation of $r = 0.45$ is significant ($p < 0.001$), as some relationship with general intelligence is expected for any complex achievement test. However, the key finding is that this correlation (0.45) is significantly lower than the correlation with the SSAT (0.78). This pattern confirms that the PABA is measuring a distinct construct—PISA-adaptive biology competence—that requires specific subject knowledge and skill, and is not simply serving as a proxy measure for general intelligence or cognitive ability. The lower, yet still significant, correlation supports the conclusion that the PABA successfully measures the unique dimensions of the intended biology framework.

Conclusions

Based on the results, the following conclusions were deduced:

1. The framework is deemed highly relevant by experts and is psychometrically sound enough to proceed with use, provided targeted revisions are made to underperforming sections.
2. The presence of defective items (like Item 36) and the marginally valid "Interpret Data" competency necessitate targeted revision to maximize the overall quality and consistency of the assessment.
3. The assessment is a dependable and precise tool for measuring PISA-adaptive biology competency, making it suitable for high-stakes decision-making.
4. The assessment successfully functions as a subject-specific measure of biology achievement, distinct from general cognitive ability, which enhances its diagnostic utility.

Recommendations

Based on the conclusions, the following are recommended:

1. The "Interpret Data" competency items must be immediately reviewed and revised to clarify their relevance and achieve the minimum required Content Validity Ratio (0.62).
2. Item 36, and any other items with negative or poor discrimination, must be deleted or thoroughly revised before finalization to ensure the integrity of the test scores.
3. After item revision and deletion, a final reliability analysis must be conducted to confirm the Cronbach's Alpha remains above the 0.80 threshold, leveraging the increase observed from the removal of flawed items.
4. Future research should focus on establishing Predictive Validity by correlating PABA scores with later academic success in higher education science courses.

Compliance with Ethical Standards

This study strictly adhered to ethical principles to protect the rights, dignity, and well-being of all human participants. The entire research protocol was designed to ensure ethical conduct from start to finish.

First, the researcher secured gatekeeper permissions before initiating any part of the study. Formal letters requesting permission to conduct the study were submitted to the Schools Division Superintendent of Mabalacat City and the principals of the selected

secondary schools. Furthermore, the complete research plan underwent a thorough review and received approval from the researcher's university ethics review committee.

The principle of informed consent was paramount. For the expert panel, a detailed consent form was provided, outlining the purpose of their participation and ensuring confidentiality. For the student participants, who were minors, a two-tiered consent process was implemented. First, informed parental consent was obtained through a take-home consent form. Second, on the day of testing, the students themselves were asked for their voluntary assent after the study was explained to them in age-appropriate language. It was made clear that participation was entirely voluntary and would not affect their class standing or grades.

To ensure confidentiality and anonymity, all data were de-identified. Students' and experts' names were not used on any research materials. Instead, alphanumeric codes were assigned to all test papers and data files. All physical documents were stored in a locked cabinet, and digital files were stored on a password-protected computer accessible only to the researcher.

Finally, the study upheld the principles of beneficence and non-maleficence (to do good and to do no harm). The potential for psychological stress from testing was minimized by creating a positive and low-pressure testing environment. It was reiterated to the students that the test was for research purposes only and the results would not impact their academic records in any way.

Acknowledgments

The researcher expresses profound gratitude to all individuals and groups who extended invaluable support in the completion of this study. With sincere appreciation, the researcher acknowledges those whose guidance, encouragement, and assistance made this academic endeavor possible.

The researcher extends heartfelt gratitude to her professor for her steadfast mentorship and invaluable guidance throughout the doctoral journey. Her expertise, dedication, and commitment to academic excellence have significantly contributed to the researcher's scholarly growth and the successful completion of this work. Sincere appreciation is likewise conveyed to the distinguished faculty members of the Graduate School and the LNU–Dagupan Thesis Committee for their insightful feedback, professional guidance, and meaningful contributions that strengthened the quality and rigor of this research.

The researcher is also grateful to classmates, colleagues, the school head, head teacher, master teachers, and the University Registrar for their encouragement, cooperation, and unwavering support throughout the research process.

Deep gratitude is extended to the researcher's beloved parents, in-laws, and siblings for their constant support and inspiration. Special thanks are given to her husband and son whose love, sacrifices, patience, and understanding served as the foundation and motivation behind this achievement.

Above all, the researcher offers her highest praise and deepest gratitude to the Divine Creator for granting wisdom, strength, guidance, and countless blessings that made the completion of this study possible.

REFERENCES

- Ahlam Lee. (2023). The importance of cultivating awareness of environmental matters in science classrooms: A cross-regional study. *Australian Journal of Environmental Education*, 39(4), 467–491. <https://doi.org/10.1017/aee.2023.22>
- Bilican Demir, S., & Yildirim, O. (2021). Indirect effect of economic, social, and cultural status on immigrant students' science performance through science dispositions: A multilevel analysis. *Education and Urban Society*, 53(3), 336–356. <https://doi.org/10.1177/0013124520928602>
- Cansiz, N., & Cansiz, M. (2019). Evaluating Turkish science curriculum with PISA scientific literacy framework. *Turkish Journal of Education*, 8(3), 217–236. <https://doi.org/10.19128/turje.522497>
- El Masri, Y. H., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Applied Measurement in Education*, 33(2), 174–188. <https://doi.org/10.1080/08957347.2020.1732384>
- Erduran, S., El Masri, Y., Cullinane, A., & Ng, Y. P. D. (2020). Assessment of practical science in high-stakes examinations: A qualitative analysis of high performing English-speaking countries. *International Journal of Science Education*, 42(9), 1544–1567. <https://doi.org/10.1080/09500693.2020.1769876>
- Guo, H., & Ercikan, K. (2020). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation*, 26(5-6), 302–327. <https://doi.org/10.1080/13803611.2021.1963941>
- Ivanova, M., Michaelides, M., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance? *Educational Research and Evaluation*, 26(5-6), 252–274. <https://doi.org/10.1080/13803611.2021.1963939>
- König, C., Khorramdel, L., Yamamoto, K., & Frey, A. (2021). The benefits of fixed item parameter calibration for parameter accuracy in small sample situations in large-scale assessments. *Educational Measurement: Issues and Practice*, 40(1), 17–27. <https://doi.org/10.1111/emip.12381>
- Organisation for Economic Co-operation and Development. (2019). *PISA 2018 assessment and analytical framework: Reading, mathematics and science*. OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- Örnek, F., Afari, E., & Alaam, S. A. (2024). Relationship between students' ICT interactions and science achievement in PISA 2018: The case of Türkiye. *Education and Information Technologies*, 29(11), 13413–13435. <https://doi.org/10.1007/s10639-023-12388-y>
- Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, 22(2), 154–184. <https://doi.org/10.1080/15305058.2022.2036161>
- Serder, M., & Jakobsson, A. (2016). Language games and meaning as used in student encounters with scientific literacy test items. *Science Education*, 100(2), 321–343. <https://doi.org/10.1002/sce.21199>
- Smith, P., Frazier, P., Lee, J., & Chang, R. (2018). Incongruence between native and test administration languages: Towards equal opportunity in international literacy

assessment. *International Journal of Testing*, 18(3), 276–296.
<https://doi.org/10.1080/15305058.2017.1407767>

Wan, Z. H., Zhan, Y., & Zhang, Y. (2024). Positive or negative? The effects of scientific inquiry on science achievement via attitudes toward science. *Science Education*, 108(1), 3–24.
<https://doi.org/10.1002/sce.21825>

APA Citation:

Nagano, A. B. (2026). DEVELOPMENT AND PSYCHOMETRIC VALIDATION OF A PISA-ADAPTIVE BIOLOGY ASSESSMENT FRAMEWORK FOR SECONDARY EDUCATION. *Ignatian International Journal for Multidisciplinary Research*, 4(9), 1237–1259. <https://doi.org/10.5281/zenodo.22702999>

analynnagano34@gmail.com