HNUE JOURNAL OF SCIENCE
Educational Sciences 2026, Volume 71, Issue 2, pp. 101-114
This paper is available online at http://hnuejs.edu.vn/es
DOI: 10.18173/2354-1075.2026-0029

DEVELOPMENT AND VALIDATION OF A SENIOR HIGH SCHOOL STATISTICS AND PROBABILITY ACHIEVEMENT TEST USING ITEM RESPONSE THEORY

Jerome L. Buhay
Department of Mathematics and Statistics, De La Salle University-Dasmariñas, Cavite, Philippines
*Corresponding author: Jerome L. Buhay, e-mail: jlbuhay@dlsud.edu.ph

Received February 24, 2026. Revised March 25, 2026. Accepted March 30, 2026.

Abstract. This study addressed the critical need for sophisticated diagnostic tools in Philippine mathematics education by developing and validating the Senior High School Statistics and Probability Achievement Test (SHSTAT). Validation in this study refers to established internal structural evidence and psychometric calibration, assessed through unidimensionality, local independence, and item-model fit. Utilizing a descriptive-developmental research design grounded in Item Response Theory (IRT), the study transcended the limitations of Classical Test Theory to provide precise measurement across the ability continuum. The instrument was administered to 1,703 Grade 11 students from public and private schools in Cavite. Results from a Modified Parallel Analysis (MPA) confirmed the instrument's unidimensionality, with all item factor loadings ($\lambda \geq 0.90$) substantially exceeding the 0.30 criterion. Comparative model fitting identified the Three-Parameter Logistic (3PL) model as the superior fit for multiple-choice data, effectively accounting for item discrimination (*a*), difficulty (*b*), and pseudo-guessing (*c*). Iterative refinement resulted in a 30-item scale that satisfied the assumption of local independence ($|Q3| < 0.20$) and exhibited excellent item-fit indices (RMSEA < 0.05). Distinct from traditional procedural assessments, item- and test-level IRT analyses indicate that the SHSTAT demonstrates high conditional measurement precision, with the Test Information Function (TIF) peaking at 17.5 around $\theta \approx 1.3$, indicating optimal precision at moderately high proficiency levels. This precision is driven by highly discriminating items, whose Item Characteristic Curves (ICCs) exhibit steep slopes and whose Item Information Functions (IIFs) show concentrated information between $0.8 < \theta < 1.8$. Consistently, the Test Characteristic Curve (TCC) displays a pronounced slope within this same range, confirming strong differentiation among higher-ability examinees. This study contributes a psychometrically robust instrument for competitive academic placement and offers educators a reliable and accurate means of assessing students' cognitive skills in Statistics and Probability, supporting data-driven instruction and curriculum refinement.

Keywords: Item Response Theory (IRT), cognitive assessment, Statistics and Probability, test development, senior high school, 3PL IRT models, Rasch model.

1. Introduction

In 2013, the Philippines overhauled its basic education system through the K–12 reform, integrating Statistics and Probability (SP) as a core subject in Senior High School (SHS). This curricular shift aimed to foster 21st-century competencies, including critical thinking, analytical problem-solving, evidence-based decision-making, and statistical literacy in today's education [1]. Despite this enhanced curriculum, a significant gap remains between intended pedagogical outcomes and actual student proficiency.

Both domestic and international studies reveal persistent pedagogical struggles in SP. Locally, SHS graduates often lack the quantitative readiness required for college, specifically in inferential reasoning and data interpretation [2], [3]. Diagnostic classroom-based assessments in regions like Pampanga show alarming performance gaps, with students averaging only 35% on foundational topics such as random variables and hypothesis testing [4]. Internationally, similar cognitive challenges have been documented. Studies across Southeast Asia and Africa frequently reveal deep-seated misconceptions about probability, randomness, and sampling distributions; students often misinterpret graphical and tabular data and struggle to transfer statistical knowledge to applied settings [5]-[7]. Large-scale assessments by the Organization for Economic Co-operation and Development (OECD) confirm that students worldwide struggle to apply statistical reasoning to authentic and real-world problems [8]. These difficulties are often exacerbated by traditional assessments that prioritize procedural rote memorization over conceptual depth [9]. Collectively, these findings underscore a critical systematic concern: the absence of psychometrically reliable, standardized assessments to evaluate high-order cognitive skills in SP.

These global and local trajectories highlight a critical need for more sophisticated diagnostic instruments. While Classical Test Theory (CTT) provides a foundation for initial item analysis, it is inherently constrained by its sample-dependency, meaning item statistics vary based on the specific group of examinees [10]. In contrast, Item Response Theory (IRT) offers a robust psychometric framework by modeling the relationship between an examinee's latent ability and individual item parameters such as difficulty, discrimination, and pseudo-guessing [11].

More importantly, IRT enables the development of assessments that provide conditional measurement precision across the entire ability continuum. Its application in large-scale international assessments such as TIMSS and PISA underscores its utility for creating equitable, scalable, and diagnostic instruments [12]. Leveraging IRT facilitates more accurate measurement, empowering educators to map learning progressions and diagnose cognitive misconceptions with high empirical clarity.

Despite the recognized advantages of IRT, its systematic application in developing standardized assessments for Senior High School Statistics and Probability (SHSP) within the Philippine context remains limited. There is a pressing need for an empirically validated instrument capable of accurately measuring students' cognitive constructs in statistical reasoning and probabilistic thinking. In accordance with the *Standards for Educational and Psychological Testing* [13], this study adopts a unitary model of validity, specifically providing internal structural evidence through the assessment of dimensionality, local independence, and item-model fit within an IRT framework.

In response, the present study aims to develop and validate the Senior High School Statistics Probability Achievement Test (SHSTAT) using an iterative IRT-based calibration framework. Specifically, the study investigates:

(1) whether the initial item pool demonstrates essential unidimensionality sufficient to support IRT modeling.

(2) the most appropriate logistic model (1PL, 2PL, or 3PL) for representing the response data and guiding item refinement.

(3) whether the retained items satisfy local independence and exhibit adequate model–data fit.

(4) how item- and test-level information functions characterize the measurement precision and effectiveness of the finalized instrument across the latent ability continuum.

Through this rigorous analysis of internal structural evidence and psychometric calibration, the study contributes a scientifically sound measurement tool that can offer evidence-based instructional insights and inform policy decisions in mathematics education. While establishing criterion-related and convergent validity remains a goal for future research, the current study provides the essential psychometric foundation by ensuring the instrument's structural integrity through IRT-based validation.

2. Content

2.1. Methodology

2.1.1. Research design

This study employed a descriptive-developmental research design to systematically develop and validate an IRT-based cognitive assessment for SP. This dual-approach design enabled the rigorous construction, interactive refinement, and psychometric evaluation of the SHSTAT, while currently providing a detailed characterization of its underlying cognitive domains and measurement properties [14]. The descriptive component provides an analytical account of the test structure and its alignment with specific cognitive strands in the curriculum. In contrast, the developmental component focuses on the technical evolution of the instrument. This process encompassed constructing an initial pool aligned with prescribed curricular competencies, verifying fundamental IRT assumptions such as unidimensionality and local independence, selecting and fitting an appropriate logistic IRT model, estimating item parameters, and conducting comprehensive item- and test-level analyses to evaluate reliability, conditional precision, and measurement effectiveness.

2.1.2. Participants

A total of 1,703 Grade 11 students from selected public (n = 1,000) and private (n = 703) senior high schools in Cavite participated in the study. All participants were enrolled in Statistical and Probability (SP) during the second semester of the 2024-2025 academic year. Participation was voluntary and governed by informed consent. The sample size significantly exceeded recommended thresholds for dichotomous IRT models, ensuring stable and accurate estimation of item parameters for 1PL, 2PL, and 3PL frameworks [15].

The selection of Cavite as a representative research site is supported by the PISA 2022 National Report (OECD, 2023) and regional achievement data from DepEd CALABARZON. The Philippines' national mean score in Mathematics (355) indicates a Level 1b proficiency; regional assessments confirm that students in Cavite exhibit performance metrics consistent with this national average [16]. Furthermore, Cavite's school density and public-private distribution (58.7% public in this study) align with the stratified nature of the national K-12 landscape, making it a defensible proxy for the initial psychometric calibration of a national-level curriculum tool like the SHSTAT.

2.1.3. Instrument

The SHSTAT is a 40-item multiple-choice assessment covering six fundamental topics in SP: Random Variables and Probability Distribution, Normal Distribution, Sampling and Sampling Distribution, Estimation, Hypothesis Testing, and Correlation and Regression Analysis (Table 1).

Items were rigorously aligned with the Grade 11 SP competencies prescribed by the Department of Education and designed to assess higher-order cognitive skills, especially problem solving, decision making, and statistical reasoning. Content validation was conducted by a panel of experts, including two senior high school SP teachers and two university statistics faculty, with iterative revisions implemented based on their expert recommendations.

Table 1. Number of items per topic

Topic	Number of items	Percentage (%)
Random Variables and Probability Distribution	6	15
Normal Distribution	8	20
Sampling and Sampling Distribution	6	15
Estimation	8	20
Hypothesis Testing	6	15
Correlation and Regression Analysis	6	15
Total	**40**	**100**

2.1.4. Data collection

The SHSTAT was administered as a proctored paper-and-pencil examination with a 90-minute time limit. Responses were dichotomously scored (1 = correct, 0 = incorrect) to facilitate IRT-based parameter estimation. Institutional permission was obtained from school principals, and informed consent was secured from students and their parents or guardians. Participation was strictly voluntary, and confidentiality and ethical standards were maintained throughout the research process.

2.1.5. Statistical analysis

All analyses were conducted using RStudio (via the 'mirt' package), supporting IRT estimation and diagnostics for 1PL, 2PL, and 3PL models. The assumption of unidimensionality for the initial 40 items was evaluated via Modified Parallel Analysis (MPA) to ensure that a single latent trait explains the item response patterns. [11], [17].

Model fit was assessed to determine the most appropriate IRT model. Competing models were compared using log-likelihood ratio tests (LRT) and information criteria, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Sample-Size Adjusted BIC (SABIC), with lower values indicating better fit [11], [16]. The selected model was used to estimate item parameters such as discrimination (a), difficulty (b), and guessing (c for 3PL) using Marginal Maximum Likelihood (MML) estimation. To ensure psychometric integrity, item calibration followed an iterative process, where items failing to meet the established thresholds were systematically removed to achieve an optimal data fit. Items were evaluated based on the following criteria: $a \geq 0.50$ for discrimination, $-3 \leq b \leq +3$ for difficulty, and $c < 0.35$ for pseudo-guessing [11], [18].

Item fit was evaluated using χ^2-based statistics and RMSEA values, with RMSEA ≤ 0.05 indicating good fit, 0.05-0.08 reasonable fit, and > 0.10 poor fit [19]. Local independence was assessed via Yen's Q3 statistic, with residual correlations $|Q3| \leq 0.2$ considered acceptable [20].

Finally, item- and test-level analyses were conducted to evaluate the psychometric quality and measurement precision of the SHSTAT. Item Characteristic Curves (ICCs) and Item Information Functions (IIFs) illustrated each item's probability of correct response and information contribution across the ability spectrum. At the test level, Test Characteristic Curves (TCCs) summarized expected total scores across ability levels, and Test Information Functions

(TIFs) evaluated overall measurement precision and reliability. These analyses ensure that the final instrument accurately differentiates students across varying levels of cognitive ability and provides a statistically robust measurement for both instructional and research purposes.

2.2. Results and discussion

2.2.1. Unidimensionality of the SHSTAT

The dimensionality analysis for the initial 40-item SHSTAT cognitive assessment, using Modified Parallel Analysis (MPA), provided compelling evidence of unidimensionality. The results showed that all item factor loadings ($\lambda \geq 0.90$) substantially exceeded the stringent criterion of 0.30. This indicates a dominant, singular latent factor underlying all item responses, signifying that the tool captures a coherent cognitive construct with high measurement precision.

Establishing unidimensionality is a fundamental prerequisite for applying IRT models, including the 1PL (Rasch), 2PL, and 3PL, which assume a single latent trait explains observed response patterns. Satisfying this assumption ensures that parameter estimates for item difficulty (b), discrimination (a), and person ability (θ) remain stable, interpretable, and statistically invariant. Furthermore, the exceptionally high factor loadings suggest that the items operate as coherent indicators of the underlying construct, consistent with findings that rigorous factor retention procedures effectively minimize extraneous variance [21].

Consequently, these findings provide strong psychometric evidence that the SHSTAT measures a single cognitive dimension, supporting its suitability for IRT-based calibration and meaning score interpretation.

2.2.2. Model selection and fit

The selection of the most appropriate Item Response Theory (IRT) model was determined through a comparative fit analysis of the 1PL, 2PL, and 3PL frameworks. The model comparison was performed on the initial 40-item pool to establish the baseline psychometric framework, with the reported degrees of freedom (df) in Table 2 reflecting the parameters of the full instrument prior to item pruning. As complexity increased, the model-data fit improved substantially. The 1PL Rasch model, which assumes uniform discrimination, yielded a log-likelihood of -42,436.932. By incorporating item-specific discrimination, the 2PL model significantly improved fit (-42,006.352). Ultimately, the 3PL model - accounting for pseudo-guessing - provided the highest log-likelihood (-41,747.591), suggesting that item difficulty (b) alone cannot account for the response patterns in the SHSTAT [22].

Table 2. Model fit comparisons for 1PL, 2PL, and 3PL IRT models

Models	Log Likelihood	AIC	SABIC	BIC	χ^2	df	p
1PL vs 2PL	-42436.93	84955.86	85048.63	85178.89	861.16	39	< 0.001
	-42006.35	84172.70	84353.72	84607.87			
2PL vs 3PL	-42006.35	84172.70	84353.72	84607.87	517.52	40	<0 .001
	-41747.59	83735.18	84006.70	84387.93			
1PL vs 3PL	-42436.93	84955.86	85048.63	85178.89	1378.68	79	<0 .001
	-41747.59	83735.18	84006.70	84387.93			

The comparative fit indices further empirical support for the 3PL model. The transition from 1PL to 2PL showed statistically significant improvement (χ^2 = 861.16, df = 39, p < .001), emphasizing the necessity of varying discrimination parameters [23]. A further significant enhancement in model fit occurred with the 3PL model (χ^2 = 517.52, df = 40, p < .001), accompanied by a corresponding decrease in Akaike Information Criterion (AIC = 83,735), Sample-Size Adjusted BIC (SABIC = 84,007), and Bayesian Information Criterion (BIC = 84,388) values. These findings justify the adoption of the 3PL model for the SHSTAT data, which is particularly suited for multiple-choice formats where low-ability examinees may correctly endorse items through pseudo-guessing [24].

2.2.3. Iterative item refinement and parameter estimation

Following the selection of the 3PL model, an iterative refinement process was employed to identify and prune problematic items, ensuring the SHSTAT meets stringent psychometric standards. Parameter estimation was conducted via Marginal Maximum Likelihood (MML).

**** Item calibration and item deletion***

The initial 40-item calibration revealed several items with suboptimal psychometric parameters. Through an iterative calibration process consisting of six calibration rounds of parameter estimation, 10 items were pruned for failing to meet the established psychometric thresholds, while the remaining 30 items demonstrated optimal model-data fit and stability (Table 3).

Item 13 was removed for negative functioning (a = -0.97), while Items 24, 35, 36, 39, and 40 were excluded for poor discrimination (a < 0.50), indicating a failure to distinguish between varying latent ability levels. Item 20 displayed an extreme difficulty parameter (b = 117.82); such non-informative items can distort the accurate estimation of a test taker's ability for the intended population [25]. Finally, items 1, 8, and 33 were eliminated as their pseudo-guessing values exceeded the 0.35 threshold, suggesting a high probability of correct endorsement through chance, which compromises measurement integrity [16].

Table 3. Iterative item calibration and deletion summary for the SHSTAT

Iteration	Item #	Discrimination (a)	Difficulty (b)	Guessing (c)	Reason for removal
1	13	-0.97	-1.95	0.277	Negative item functioning
	20	0.024	117.815	0.322	Extreme difficulty
2	1	1.529	0.916	0.39	High guessing probability
	8	2.857	1.794	0.451	High guessing probability
3	24	0.343	2.308	0.008	Poor discrimination
	35	0.33	2.141	0.007	Poor discrimination
4	36	0.424	1.708	0.006	Poor discrimination
	39	0.454	1.664	0.031	Poor discrimination
5	33	3.318	1.509	0.351	High guessing probability
	40	0.47	1.851	0.002	Poor discrimination
6					Final Model

The distribution of items per topic for the final 30-item SHSTAT is presented in Table 4. The final 30-item version of the instrument reflects a distribution rigorously aligned with the instructional weighting of the K-12 SP curriculum. While Correlation and Regression Analysis comprise a smaller subset (n = 2), these items were retained during the 3PL calibration due to their exceptional discriminating power and significant contribution to the overall Test Information Function (TIF). This approach ensures that the instrument preserves comprehensive content coverage while prioritizing the psychometric rigor necessary for a stable, interpretable, and valid internal structure.

Table 4. Distribution of the final 30-item SHSTAT by topic

Topic	Number of items	Percentage (%)
Random Variables and Probability Distribution	5	16.7
Normal Distribution	6	20.0
Sampling and Sampling Distribution	5	16.7
Estimation	7	23.3
Hypothesis Testing	5	16.7
Correlation and Regression Analysis	2	6.7
Total	30	100.0

**** Final parameter properties (30-item model)***

Following the iterative refinement and calibration of the final 30-item SHSTAT, the resulting parameter estimates (Table 4) characterize a high-quality, psychometrically sound assessment instrument. The discrimination parameters (a) ranged from 0.838 to 3.369. Specifically, 9 items exhibited moderate discrimination ($0.5 < a < 1.5$), 17 items demonstrated high discrimination ($1.5 < a < 2.5$), while five items, such as Items 14, 15, 17, 28, and 34, exhibited very high discrimination ($a > 2.5$). High a-values are critical for precise ability estimation, as they indicate that the items are highly sensitive to small differences in the latent trait [11]. Thus, this suggests that the SHSTAT can accurately differentiate between examinees of similar proficiency levels, particularly in competitive academic environments where precise classification is vital.

The difficulty parameters (b) ranged from 0.237 to 2.085, with a mean difficulty concentrated in the moderate-to-high range. While Items 18 ($b = 0.237$) and 19 ($b = 0.787$) provide measurement for examinees with slightly above-average ability, the majority of the items (e.g., Items 7, 29, and 31) are targeted toward high-proficiency individuals ($\theta \approx 2.0$). This distribution enables the test to measure examinees across a broad ability spectrum while offering optimal targeting of higher-ability individuals. Consequently, the SHSTAT is optimized for identifying top-tier performance and distinguishing among high-ability candidates rather than diagnosing severe deficits at the lower end of the ability scale.

Finally, the pseudo-guessing parameters (c) ranged from 0.002 to 0.308. Although some items approached the upper limit, all remained within the permissible $c < 0.35$ threshold, consistent with recommended practice [18]. Overall, the relatively low c-values suggest that the distractors within the multiple-choice format are functioning effectively, thereby minimizing the impact of chance on the scores and preserving the integrity of the latent trait measurement. Nevertheless, the guessing parameters for Items 22, 23, 31, and 34, while psychometrically acceptable, are close to the theoretical chance level of 0.25. This proximity suggests that certain distractors may be insufficiently plausible for lower-ability examinees. Accordingly, these items

will undergo targeted qualitative review and distractor refinement to better engage common misconceptions, thereby improving measurement precision across a wide range of the ability scale.

The findings suggest that the 30-item SHSTAT provides a highly precise measurement of cognitive proficiency by filtering out non-performing items and effectively controlling for the influence of random guessing. Its alignment with moderate-to-high difficulty levels makes it a psychometrically defensible instrument for identifying top-tier students in high-stakes academic selection. Consequently, the refined assessment offers a robust and stable metric, ensuring that observed scores are a valid reflection of a student's true latent ability.

Table 4. Parameter estimates of the 3PL IRT model for 30 items

Items	Parameters			Items	Parameters		
	a	***b***	***c***		***A***	***b***	***c***
Item 2	1.326	1.171	0.299	Item 19	2.159	0.787	0.214
Item 3	1.408	1.155	0.230	Item 21	0.838	1.705	0.002
Item 4	1.573	1.048	0.255	Item 22	2.347	1.777	0.276
Item 5	1.468	1.333	0.197	Item 23	2.285	1.717	0.280
Item 6	2.427	1.224	0.226	Item 25	1.664	1.132	0.138
Item 7	1.568	1.933	0.239	Item 26	1.564	1.101	0.182
Item 9	2.025	1.004	0.240	Item 27	1.283	1.727	0.143
Item 10	2.355	1.272	0.239	Item 28	3.369	1.396	0.243
Item 11	1.547	1.126	0.263	Item 29	1.107	2.085	0.232
Item 12	1.430	1.272	0.091	Item 30	1.791	1.390	0.149
Item 14	3.216	1.012	0.238	Item 31	1.680	1.940	0.308
Item 15	2.968	1.082	0.241	Item 32	2.081	1.406	0.210
Item 16	1.916	1.192	0.117	Item 34	2.587	1.337	0.294
Item 17	2.828	1.019	0.179	Item 37	1.138	1.145	0.041
Item 18	0.841	0.237	0.002	Item 38	1.598	1.301	0.137

2.2.4. Item fit and local independence

To confirm the structural adequacy of the 3PL IRT model, item fit was rigorously evaluated using both chi-square (χ^2) statistics and the RMSEA. While the χ^2 statistic is a traditional measure of exact model fit, it is notoriously sensitive to large sample sizes, often yielding statistically significant results ($p < .05$) for trivial deviations that do not impact practical measurement [26]. Indeed, as shown in Table 5, several items yielded significant p-values; however, a purely frequentist interpretation would be misleading in this context. To mitigate this, RMSEA was employed as a parsimony-adjusted index to assess the magnitude of misfit. All 30 items demonstrated excellent fit to the 3PL model with RMSEA values below 0.05, indicating a high degree of congruence between the observed response patterns and the model's predictions [27]. Item 7 exhibited a minor deviation with an RMSEA of 0.054; however, this remains within the acceptable threshold of 0.08, indicating negligible misfit [28].

Table 5. Item fit statistics for the 3PL IRT model

Item	χ2	df	RMSEA	p	Item	χ2	df	RMSEA	p
Item 2	18.419	21	0.000	0.622	Item 19	17.942	19	0.000	0.526
Item 3	25.696	21	0.011	0.218	Item 21	53.673	23	0.028	0.000
Item 4	23.555	20	0.010	0.262	Item 22	27.872	21	0.014	0.144
Item 5	16.315	21	0.000	0.752	Item 23	33.288	21	0.019	0.043
Item 6	32.372	20	0.019	0.039	Item 25	18.699	20	0.000	0.541
Item 7	131.097	22	0.054	0.000	Item 26	29.457	20	0.017	0.079
Item 9	25.464	19	0.014	0.146	Item 27	27.162	22	0.012	0.205
Item 10	18.519	20	0.000	0.553	Item 28	37.210	20	0.022	0.011
Item 11	46.060	20	0.028	0.001	Item 29	30.549	23	0.014	0.134
Item 12	34.800	21	0.020	0.030	Item 30	49.089	21	0.028	0.000
Item 14	27.523	18	0.018	0.070	Item 31	31.458	22	0.016	0.087
Item 15	22.579	19	0.011	0.256	Item 32	25.709	20	0.013	0.176
Item 16	22.191	20	0.008	0.330	Item 34	38.309	20	0.023	0.008
Item 17	21.903	19	0.009	0.289	Item 37	36.965	22	0.020	0.024
Item 18	87.200	21	0.043	0.000	Item 38	34.812	21	0.020	0.030

Complementing the item fit analysis, the Local Independence (LD) assumption was examined using Yen's Q3 statistic. Local independence requires that, once the primary latent trait (θ) is accounted for, the residual responses to any pair of items are statistically uncorrelated. Violations of this assumption often signify item redundancy or the presence of an unintended secondary dimension, which can lead to inflated reliability and biased parameter estimates.

Table 6. Summary of Yen's Q3 test of local independence

Statistic	Q3	Adjusted Q3
Minimum	-0.107	-0.087
Maximum	0.095	0.115
Mean	-0.020	

The analysis of the SHSTAT yielded residual correlations ranging from -0.107 to 0.095 (Table 6), with all |Q3| values remaining well below the commonly cited threshold of 0.20. Examination of the Yen's Q3 matrix indicated that the highest residual correlation occurred between Items 23 and 24 (r = 0.095), while the lowest was observed between Items 12 and 16 (r = -0.107). The uniformly low magnitude of these correlations suggests a negligible degree of local dependence among items.

This pattern provides structural internal evidence that the SHSTAT items function independently and do not exhibit meaningful redundancy. The absence of substantial local dependence supports the assumption that the 30-item instrument is primarily measuring a single latent construct, thereby contributing to the stability and interpretability of the resulting ability estimates within the studied context.

2.2.5. Item and test analyses of the effectiveness and precision of the SHSTAT

**** Item Characteristic Curves (ICCs) and Item Information Functions (IIFs) analyses***

To evaluate the psychometric quality and functional utility of the SHSTAT, item-level analyses were conducted using the principles of the 3PL IRT model. This framework facilitates a detailed examination of item behavior through Item Characteristic Curves **(**ICCs) and measurement precision through (Item Information Functions) IIFs. The 30 retained items were categorized into moderate (0.65–1.34), high (1.35–1.69), and very high (≥1.70) discrimination groups, allowing for a comparative assessment of how varying discrimination levels distinguish examinees across the ability continuum (θ) [23], [28].

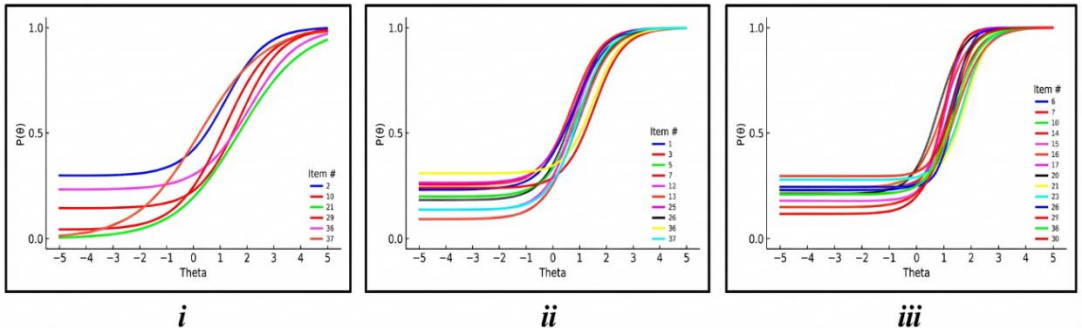

Figure 1. Item characteristic curves of moderate (i), high (ii) and very high (iii) discrimination items

The ICC analysis (Figure 1) reveals the combinations among difficulty (b), discrimination (a), and pseudo-guessing (c). Items with moderate discrimination offer stable differentiation across broad proficiency levels, with difficulties ranging from $b = 0.24$ (Item 18) to $b > 1.7$ (Items 21, 27, and 29). However, elevated baselines in Items 2 and 29 ($c > 0.22$) suggest that low-ability examinees may achieve correct endorsement through chance (De Ayala, 2022). As discrimination increases, gradients steepen significantly, reflecting enhanced measurement precision. Very high discrimination items (e.g., 14, 15, 17, and 28) demonstrate extreme sensitivity to subtle ability differences within the $0.8 < \theta < 1.8$ range. Despite their high accuracy, several items (22, 23, and 34) exhibit c-values near 0.25, which can destabilize parameter estimation if not carefully monitored.

The IIF analysis (Figure 2) confirms that precision is highest among very discriminating items. While items with moderate discrimination provide broad but low information (0.18 - 0.31), very high discrimination items yield tall, narrow peaks. Notably, Item 28 provides maximum precision (1.78) at $\theta \approx 1.50$, illustrating that while high discrimination maximizes accuracy, the informational contribution is tightly concentrated near the item's difficulty parameter [23].

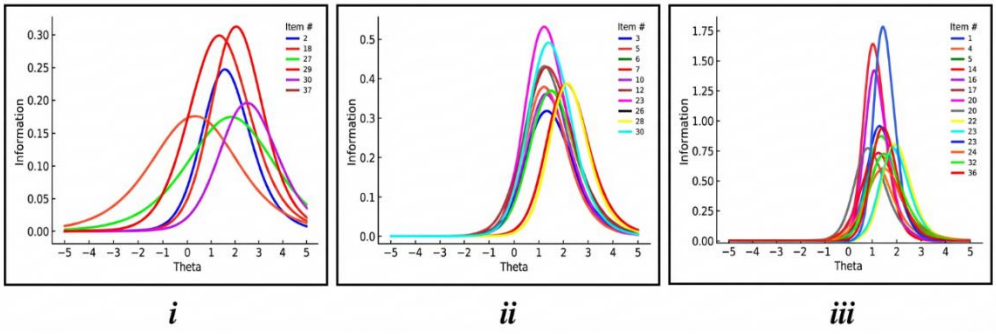

Figure 2. Item information functions of moderate (i), high (ii) and very high (iii) discrimination items

The analysis indicates that SHSTAT measurement precision is concentrated in the middle-to-upper ability range. The influence of pseudo-guessing ($0.21 \leq c \leq 0.30$) reduces measurement precision for low-ability examinees and can bias latent trait estimates. Consequently, the SHSTAT is a robust instrument for identifying high-proficiency individuals but lacks diagnostic sensitivity for lower-ability groups. Future refinements should prioritize distractor revision for items with high guessing parameters and the addition of low-difficulty items ($b < 0$) to ensure a more balanced and inclusive proficiency measure.

**** Test Characteristic Curve (TCC) and Test Information Functions (TIF) analyses***

Following the item-level calibration, test-level properties were evaluated using the Test Characteristic Curve (TCC) and Test Information Function (TIF) to determine the instrument's overall precision and expected scoring patterns. The TCC (Figure 3) illustrates the relationship between the latent trait (θ) and the expected total score, following a prominent S-shape with the most significant slope occurring between $\theta = 0.5$ and $\theta = 2.5$. This confirms that the SHSTAT is most sensitive to changes in ability within the average to high-proficiency range. Notably, the lower asymptote of the TCC settles near five points rather than zero, an elevation resulting from the aggregate pseudo-guessing parameters (c) identified in the 30-item pool. This suggests that random success exerts a measurable influence on the scores of low-ability examinees, thereby compressing the scale at the lower end.

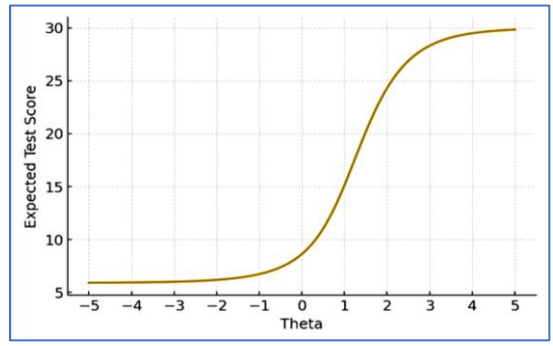

Figure 3. Test characteristics curve (TCC) of the SHSTAT assessment

Complementing this, the Test Information Function (TIF) (Figure 4) depicts the distribution of measurement information across the ability continuum, displaying a bell-shaped distribution that peaks at $\theta \approx 1.3$ with a maximum measurement information value of 17.5. This peak marks the point of minimum standard error and maximum reliability. The SHSTAT maintains a stable information level between $\theta = 0.5$ and $\theta = 1.5$, indicating that the test provides highly dependable measurement for the upper-average proficiency range. However, measurement precision declines quickly outside this range, especially at lower ability levels, which shows that the test functions better as a selection tool than as a basic diagnostic instrument for identifying learning deficits.

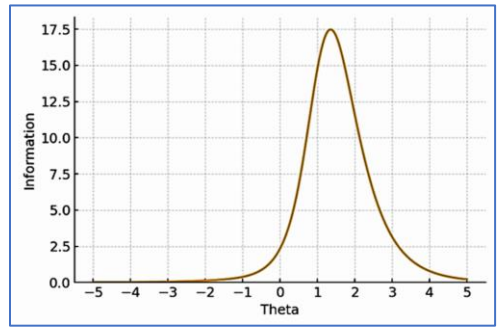

Figure 4. Test information function (TIF) of the SHSTAT assessment

Collectively, the TCC and TIF data indicate that the SHSTAT is a structurally robust instrument for competitive academic placement. While the instrument provides excellent differentiation among high-ability candidates, the concentration of information in the above-average proficiency range constrains its effectiveness for lower-performing student groups. To enhance the instrument's internal structural validity, future refinements should prioritize the inclusion of low-difficulty items ($b < 0$) and the strengthening of distractor quality to mitigate pseudo-guessing effects (c). Such adjustments would broaden the range of reliable measurement, transforming the SHSTAT from a specialized selection tool into a more comprehensive proficiency assessment.

3. Conclusions

The SHSTAT demonstrates strong internal structural validity. Grounded in Item Response Theory (IRT), the instrument demonstrated satisfactory adherence to the core assumptions of unidimensionality and Local Independence. (LD) Comparative model-fitting analyses indicated that the Three-Parameter Logistic (3PL) model provided the best empirical representation of the data, capturing the interaction among item discrimination (*a*), difficulty (*b*), and pseudo-guessing (*c*).

At the item level, the SHSTAT exhibits generally high discrimination and difficulty parameters calibrated toward average-to-high ability examinees. Although the instrument demonstrates strong model fit and high measurement precision for distinguishing higher proficiency levels, analyses revealed reduced sensitivity at the lower end of the ability continuum, partly attributable to the influence of guessing. Overall, the results provide structural internal evidence supporting the psychometric adequacy of the 30-item SHSTAT for estimating student ability along a single latent dimension, as defined by the IRT framework.

Moving forward, future studies should focus on establishing criterion-related validity by correlating SHSTAT scores with students' final grades in Statistics and Probability to further strengthen its diagnostic utility. Future research should also prioritize the transition of the SHSTAT into a Computerized Adaptive Testing (CAT) framework, utilizing the established 3PL parameters to dynamically tailor item difficulty and maximize precision across all proficiency levels.

While the SHSTAT may be used for instructional and research purposes based on its established psychometric calibration and structural internal evidence, Differential Item Functioning (DIF) analyses are still required to confirm measurement invariance. This is essential to ensure equitable score interpretation across subgroups, such as gender and SHS curriculum strands (e.g., STEM, ABM, HUMSS), before large-scale implementation. Furthermore, longitudinal studies should explore the expansion of the item bank to include lower-difficulty items ($b < 0$) with refined distractor quality. Replicating this IRT-based methodology in other core disciplines, such as Mathematics and Science, would support the development of a standardized, psychometrically coherent assessment framework for Senior High School.

Acknowledgments. This research was partly presented at the 4th International Conference on Innovation in Learning Instruction and Teacher Education (ILITE4), Hanoi, 13-14 December 2025.

REFERENCES

[1] Department of Education, (2013). *K to 12 Basic Education Curriculum: Senior high school core subject—Statistics and probability (Curriculum guide)*. Republic of the Philippines. https://www.deped.gov.ph/wp-content/uploads/2022/02/SHS-Core_Statistics-and-Probability-CG.pdf.

[2] Mamba M, Tamayao A & Vecaldo R, (2020). College readiness of Filipino K to 12 graduates: Insights from a criterion-referenced test. *International Journal of Education and Practice,* 8(4), 625-637. https://doi.org/10.18488/journal.61.2020.84.625.637.

[3] Paat FMG, (2023). School motivation, learning strategies and college readiness of senior high school graduates in the Philippines. *Journal for Educators, Teachers and Trainers,* 14(3), 1-8. https://jett.labosfor.com/index.php/jett/article/view/1620.

[4] Calma JD, Salvador IGO & Supan AM, (2022). Knowledge and attitude toward statistics and probability of senior high school students. *Asia Pacific Journal of Educational Perspective,* 9(1). https://research.lpubatangas.edu.ph/wp-content/uploads/2022/09/3-APJEP-2022-38-Calma-et-al..pdf.

[5] Tan SH & Vighnarajah V, (2025). Exploring students' misconceptions in the probability topic of Form 4 mathematics. *Malaysian Journal of Social Sciences and Humanities,* 9(S1), 251-262. https://doi.org/10.47405/mjssh.v9iS1.3005.

[6] Sari DP, Suryadi D & Dasari G, (2024). Learning obstacle of probability learning based on the probabilistic thinking level. *Journal on Mathematics Education,* 15(1), 207-228. https://doi.org/10.22342/jme.v15i1.

[7] Makonye JP & Fakude J, (2016). A study of errors and misconceptions in the learning of addition and subtraction of directed numbers in Grade 8. *SAGE Open,* 6(4), 1-10. https://doi.org/10.1177/2158244016671375.

[8] Organisation for Economic Co-operation and Development, (2023). *PISA 2022 results (Volume I): The state of learning and equity in education.* OECD Publishing. https://doi.org/10.1787/53f23881-en.

[9] Batanero C & Álvarez-Arroyo R, (2024). Teaching and learning of probability. *ZDM–Mathematics Education,* 56, 5-17. https://doi.org/10.1007/s11858-023-01511-5.

[10] Mamolo LA, (2021). Development of an achievement test to measure students' competency in general mathematics. *Anatolian Journal of Education,* 6(1), 79-90. https://doi.org/10.29333/aje.2021.616a.

[11] de Ayala RJ, (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.

[12] Bezirhan U & von Davier M, (2024). *TIMSS achievement scaling methodology: Item response theory and population models.* In von Davier M, Fishbein B & Kennedy AM (Eds.), "TIMSS 2023 technical report: Methods and procedures". International Association for the Evaluation of Educational Achievement (IEA).

[13] American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

[14] Calderon JF & Gonzales EC, (2019). *Methods of research and thesis writing.* MG Reprographics Supply & Services Inc.

[15] OECD, (2023). *PISA 2022 Results* (Volume I): The State of Learning and Equity in Education. PISA, OECD Publishing, Paris. https://doi.org/10.1787/53f2383c-en.

[16] DeMars C, (2010). *Item response theory.* Oxford University Press.

[17] Glorfeld LW, (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement,* 55(3), 377-393. https://doi.org/10.1177/0013164495055003002.

[18] Baker FB & Kim SH, (2017). *The basics of item response theory using R.* Springer. https://doi.org/10.1007/978-3-319-54205-8.

[19] Browne MW & Cudeck R, (1993). *Alternative ways of assessing model fit*. In Bollen KA & Long JS (Eds.), "Testing structural equation models", p. 136-162. Sage Publications.

[20] Christensen KB, Makransky G & Horton M, (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement,* 41(3), 178-194. https://doi.org/10.1177/0146621616677520.

[21] Lim H & Jahng S, (2019). Scale linking for the testlet item response theory model. *Educational and Psychological Measurement,* 79(6), 1081-1106. https://doi.org/10.1177/0013164419844287.

[22] Kang T & Cohen AS, (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement,* 31(4), 331-358. https://doi.org/10.1177/0146621606292213.

[23] Embretson SE & Reise SP, (2025). *Item response theory: Foundations for psychologists and social scientists* (2nd ed.). Routledge. https://doi.org/10.4324/9781315726557.

[24] Gyamfi A & Acquaye R, (2023). Parameters and models of item response theory (IRT): A review of literature. *Acta Educationis Generalis,* 13(3), 68-78. https://doi.org/10.2478/atd-2023-0022.

[25] Zhang S, Wang W & Tao J, (2018). Estimating the 3PL model parameters with the maximum likelihood method and the Bayesian method. *Journal of Applied Statistics,* 45(12), 2244-2261. https://doi.org/10.1080/02664763.2017.1414163.

[26] Orlando M & Thissen D, (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement,* 24(1), 50-64. https://doi.org/10.1177/01466216000241003.

[27] MacCallum RC, Browne MW & Sugawara HM, (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods,* 1(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130.

[28] Baker FB, (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. https://eric.ed.gov/?id=ED458219.