P-ISSN: 2723-8156
e-ISSN: 2723-8164
DECEMBER 2026

U-Teach: Journal Education of Young Physics Teacher 7 (1) (2026) 99-110
https://jsr.unuha.ac.id/index.php/U-Teach

Beyond Gameplay: Item-Level Diagnostics and Psychometric Rigor of Resonance Odyssey Stealth Assessment

Nur Endah Susilowati[1*], Fitria Arifiyanti[1], Putra Habib Dhitareka[1], Muhammad Fadhlan Aprianto[1], Aliya Dewi Rahmawati[1]

[1]Faculty of Mathematics and Natural Sciences Education, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung, West Java 40154, Indonesia

*Corresponding author: nurendahsusilowati@upi.edu

Article History:
Received: October 10, 2026
Revised: October 21, 2026
Accepted: October 26, 2026
Published: Desember 01, 2026

Keywords: *game-based assessment, item-level diagnostics, psychometric rigor, rasch model, stealth assessment*

Abstract: Stealth assessment embeds educational measurement directly into gameplay to evaluate student competencies without disrupting the learning flow. However, ensuring the psychometric rigor of game-based instruments at the item level remains a crucial challenge. This study evaluated the psychometric properties and item-level diagnostics of a 30-item stealth assessment integrated into the educational board game *Resonance Odyssey*. Data were collected from 129 participants (N = 129) and analyzed using the Rasch measurement model via Winsteps 5.2.5.1. The analysis confirmed strong construct validity and unidimensionality, with 36.8% of the raw variance explained by the measure and a first residual contrast eigenvalue of 3.03 (6.4%). High internal consistency and item hierarchy were demonstrated by excellent item reliability (R_i = 0.95, separation = 4.60) and good person reliability (R_p = 0.87, separation = 2.61). Wright map targeting showed an optimal alignment between student ability and task difficulty, indicated by a mean person measure of +0.03 logit. Local independence was fully satisfied, with maximum standardized residual correlation reaching only r = 0.47. Detailed item-level diagnostics identified four misfitting items (Q22, Q25, Q26, and Q30) characterized by high outfit mean square values (MNSQ > 3.0) and unexpected option endorsements by high-ability students. Overall, the findings demonstrate that while *Resonance Odyssey* provides a psychometrically rigorous framework for non-intrusive skill measurement, targeted revision of problematic distractors is essential to optimize its diagnostic precision.

INTRODUCTION

"When fear takes over the testing room, a physics score measures panic rather than potential". Teachers administer tests on specific subject matter to evaluate students' abilities (Adesope et al., 2017; Prieto et al., 2021; Skinner, 2020; Yan et al., 2019). This is also the case in physics; however, field observations frequently reveal that despite studying repeatedly, students consistently obtain unsatisfactory scores. But is this because students are truly unable to solve the problems, or because they are already overwhelmed by fear before even starting?

This fundamental question highlights the presence of physics test anxiety as a pervasive affective barrier that severely undermines the validity of educational measurements (Deng & Liu, 2025). Conventional testing formats typically present assessment as a discrete, high-stakes evaluative event. Under these tense

testing conditions, the perceived threat of failure triggers intrusive thoughts and emotional distress, which consume working memory capacity and frequently lead to cognitive "blanking" during exams. Consequently, a student's low performance on a conventional physics test often reflects evaluative anxiety rather than a genuine lack of scientific reasoning competence (Deng & Liu, 2025). To dismantle this psychological barrier, stealth assessment has emerged as a compelling solution by embedding measurement invisibly within interactive learning environments, such as games. By continuously capturing evidence of student competencies through gameplay actions without disrupting their flow state, stealth assessment transforms testing into a low-threat mission, effectively mitigating test anxiety and freeing cognitive resources for complex scientific reasoning.

Building on this anxiety-reducing framework, *Resonance Odyssey* was developed as a game-based stealth assessment to evaluate scientific reasoning in sound-wave physics. Grounded in Evidence-Centered Design (ECD) and Mayer's Cognitive Theory of Multimedia Learning (CTML), the game leverages interactive mechanics and cultural contexts to create an engaging, low-stress evaluation climate (Melander Bowden & Aarsand, 2020; Shute & Rahimi, 2021). However, while stealth assessment successfully alleviates affective interference, ensuring its diagnostic validity requires rigorous psychometric scrutiny at the item level. Mitigating test anxiety is only half the battle; if the embedded assessment tasks contain misfitting items, ambiguous prompts, or over-keying distractors, the assessment risks introducing new sources of measurement error (Fernanda & Hidayah, 2020; Soeharto & Csapó, 2022) .

Evaluating the psychometric integrity of an anxiety-reducing stealth assessment necessitates moving beyond Classical Test Theory (CTT) toward modern item response frameworks. CTT metrics rely heavily on sample-dependent total scores, which cannot isolate item difficulty from student ability nor detect subtle item-level anomalies. In contrast, the Rasch measurement model transforms ordinal response data into linear logit scales, offering a mathematically sound framework to examine construct unidimensionality, item difficulty hierarchies, and local independence. Crucially, Rasch fit statistics (Infit and Outfit Mean Squares) serve as vital diagnostic tools to identify misfitting items where students exhibit unexpected response patterns, such as high-ability students falling into specific distractor traps despite feeling unpressured during gameplay.

While recent literature highlights the affective benefits of game-based learning in reducing test anxiety (Ahmed et al., 2022; Bonefont et al., 2022; Dumile & Boateng, 2026; Liu et al., 2026) , empirical studies providing comprehensive, item-level psychometric diagnostics of board game-based stealth assessments remain remarkably sparse (Chen et al., 2020; Lu et al., 2025). Existing research predominantly focuses on macro-level outcomes, such as overall score improvements or anxiety reduction, while leaving the micro-level psychometric health of individual game missions unexamined. Without rigorous item diagnostics, flawed task architectures or poorly functioning distractors can remain hidden behind high player engagement, compromising the accuracy of the inferences drawn about students' true reasoning abilities.

To address this gap, the present study conducts an item-level psychometric evaluation of the 30-item stealth assessment integrated within the *Resonance Odyssey* board game using the Rasch measurement model. By analyzing response data from 129 high school

students, this study examines how effectively the game-based tasks maintain psychometric rigor while operating within a low-anxiety assessment environment. Specifically, the following research questions guide this investigation:

1. **RQ1:** To what extent does the *Resonance Odyssey* stealth assessment satisfy the fundamental Rasch measurement assumptions of unidimensionality, local independence, and person-item targeting?
2. **RQ2:** What are the internal consistency and structural reliability properties of the assessment in terms of Rasch person/item separation indices?
3. **RQ3:** Which specific game-based items exhibit structural misfits, and what do option-level diagnostic dynamics reveal about distractor functionality in an interactive stealth assessment environment?

METHODS

Research Design

This study employed a quantitative psychometric evaluation design utilizing the Rasch measurement model (Joint Maximum Likelihood Estimation / JMLE) to evaluate the structural integrity, construct validity, and item-level diagnostics of a game-based stealth assessment (Bond, 2015; Fernanda & Hidayah, 2020; Linacre, 2013; You et al., 2018). Unlike Classical Test Theory (CTT) which relies on raw sum scores, the Rasch model transforms ordinal item responses into linear, interval-level logit measures, enabling independent calibration of item difficulties and person abilities (Fernanda & Hidayah, 2020) . The overall research procedure followed a four-stage methodological framework, as illustrated in Figure 1.

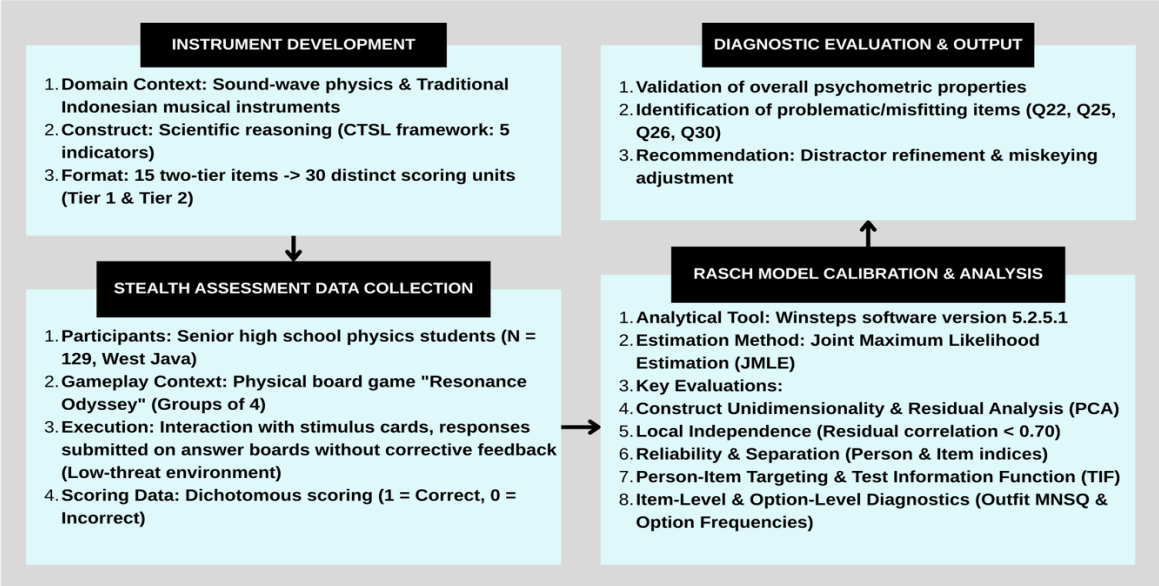


Figure 1. Methodological framework of the study, illustrating the four-stage process

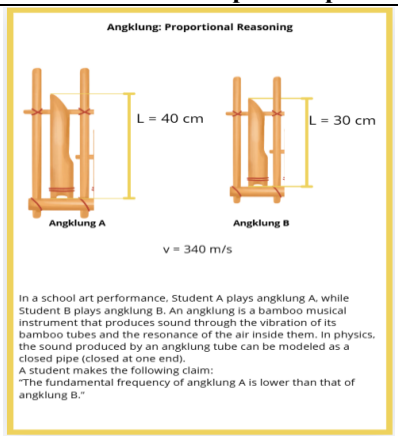
Participants

The calibration sample comprised 129 senior high school students (N = 129) enrolled in the physics track across senior high schools in West Java, Indonesia, who had completed the sound-wave physics unit. In Rasch measurement, a sample size of 129 exceeds the threshold required to achieve stable item difficulty calibrations within ± 50 logits at a 99% confidence interval for dichotomous items, ensuring sufficient statistical power for item-level diagnostics.

Instrument Architecture and Game Mechanics

The assessment was embedded within the physical board game *Resonance Odyssey*, designed around sound-wave physics concepts and contextualized using traditional Indonesian musical instruments. The instrument architecture follows the Evidence-Centered Design (ECD) framework (see **Figure 2**), incorporating a Competency Model targeting scientific reasoning (CTSL framework: proportional, control of variables, probabilistic, correlational, and hypothetical-deductive reasoning), a Task Model based on province-level challenges, an Assembly Model structuring gameplay progression, and an Evidence Model scoring observable responses. The instrument comprised 15 two-tier items, yielding 30 distinct scoring units for Rasch calibration (Tier 1 for content reasoning and Tier 2 for justification), scored dichotomously. The sample of the instruments is displayed in Table 1.

Table 1. Representative Two-Tier Stealth Assessment Item (Angklung Context)

Dimension	Item Content & Response Options	Analysis
Proportional Reasoning	 Stimulus	Establishes the physical context in which students must reason about the proportional relationship between air-column length and fundamental frequency.
Stimulus Claim		Provides a claim that students must evaluate using quantitative comparisons and conceptual understanding of the relationship between tube length and frequency.
Tier 1: Content Reasoning	"The sound frequency of angklung A is smaller than angklung B" The ratio of the sound frequency of angklung A and B ($f_a : f_b$) is … a. 3:4 b. 4:3 c. 1:1 d. 2:1 e. 1:2	Quantitative Proportional Execution: Requires students to apply the inverse proportional relationship between air-column length and fundamental frequency in a closed-pipe model. Students must calculate the inverse ratio. This assesses their ability to compare quantities through inverse proportionality rather than direct proportionality.
Tier 2: Justification	The student's statement is … A. correct, because the fundamental frequency is inversely proportional to the length of the air column B. correct, because the length of the tube determines the sound amplitude C. incorrect, because the frequency is directly proportional to the length of the air column D. incorrect, because the number of tubes determines the frequency E. incorrect, because the mass of the bamboo determines the frequency Answer : a	Causal and Relational Explanation: Evaluates whether students understand the physical principle underlying the calculated ratio rather than merely substituting values into a formula. The distractors represent potential misconceptions, including confusing frequency with amplitude, assuming direct proportionality, and attributing frequency differences to irrelevant physical variables.

Dimension	Item Content & Response Options	Analysis
Integrated Two-Tier Assessment	Students select one answer in Tier 1 and one justification in Tier 2. The correct combination is Tier 1: A + Tier 2: A.	Combining the two tiers distinguishes students who can calculate the correct ratio and explain the underlying physical relationship from those who arrive at the correct answer through guessing, memorization, or incomplete conceptual understanding.

As illustrated in Table 1, this item design demonstrates how the instrument effectively captures students' proportional reasoning through a contextualized stealth assessment framework. During gameplay, students progressed in groups of four through provincial challenges, interacted with question cards featuring physics stimuli, and submitted responses on individual answer boards under stealth assessment principles without corrective feedback (**Figure 2**).

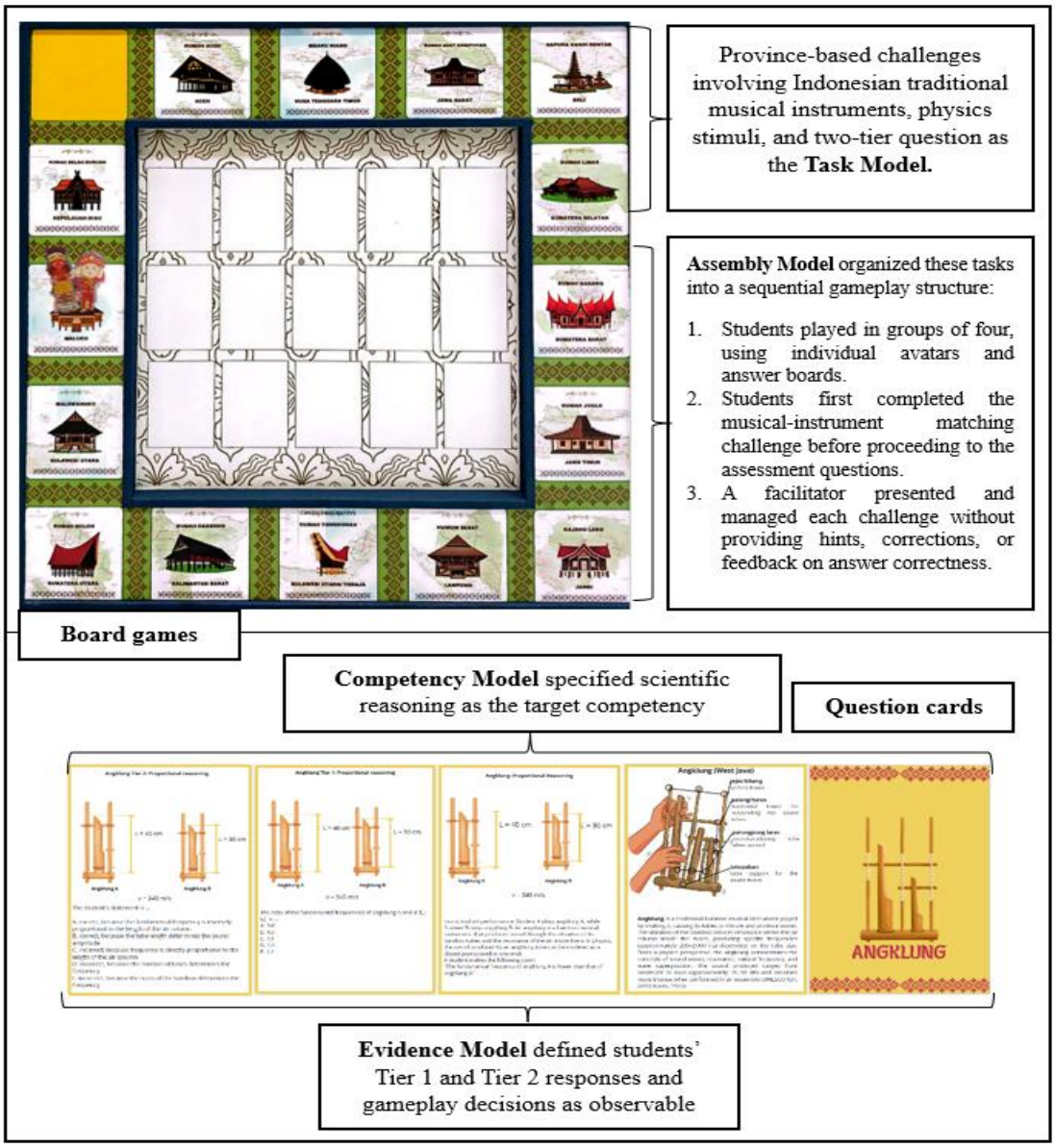


Figure 2. Operational architecture of the Resonance Odyssey stealth assessment framework, incorporating Evidence-Centered Design (ECD) components′

Based on Figure 2, each two-tier item was divided into two distinct scoring units (Tier 1 for content reasoning and Tier 2 for underlying justification), yielding 30 items ($N_{Item} = 30$) for Rasch calibration. Items were scored dichotomously (1 = correct, 0 = incorrect). During gameplay, students progressed through provincial challenges, engaged with physical cards displaying musical instrument stimuli, and submitted responses on individual answer boards under stealth assessment principles without feedback regarding correctness.

Psychometric Criteria and Data Analysis

Item response data were analyzed using Winsteps software version 5.2.5.1. The psychometric evaluation adhered to standard Rasch diagnostic criteria:

1. **Construct Validity and Unidimensionality:** Evaluated using Principal Component Analysis (PCA) of standardized residuals. Unidimensionality was satisfied if raw variance explained by measures exceeded 20% and the unexplained variance in the first contrast had an eigenvalue < 3.0 (< 15%).
2. **Local Independence:** Assessed via the largest standardized residual correlations between item pairs. Local independence was maintained if inter-item residual correlations remained below 0.70.
3. **Reliability and Separation:** Evaluated using Person and Item Reliability (R_p, $R_i \geq 0.80$) and Separation Indices (> 2.0), indicating the instrument's capacity to differentiate ability strata and establish a stable difficulty hierarchy.
4. **Item Fit Statistics:** Item quality was evaluated using Outfit Mean Square (MNSQ = 0.5 to 1.5), Outfit Z-standardized (ZSTD = -2.0 to +2.0), and Point-Measure Correlation (r_{pt}> 0.0). Items violating these parameters were classified as misfitting.
5. **Option-Level Diagnostics:** For misfitting items, distractor efficiency was examined using the Winsteps Option Frequencies Table to analyze response distributions and mean ability logits across key choices and distractors.

RESULTS & DISCUSSION

Results

Fundamental Measurement Assumptions: Unidimensionality, Local Independence, and Targeting (RQ1)

To address **RQ1**, the dataset derived from 129 respondents (N = 129) across 30 measurement units (N_{item} = 30) was evaluated against the primary assumptions of the Rasch measurement model: construct unidimensionality, local independence, and person-item targeting.

1. Construct Unidimensionality

Construct validity was examined through a Principal Component Analysis (PCA) of standardized residuals. As presented in Table 2, the raw variance explained by the Rasch model measures reached 36.8% (comprising 23.0% person variance and 13.8% item variance), successfully exceeding the 20.0% threshold required for unidimensionality. Furthermore, the unexplained variance in the first contrast accounted for an eigenvalue of 3.03 (6.4%), which remains well within the acceptable upper limit of 15.0% (eigenvalue < 3.0). These statistics confirm that the *Resonance Odyssey* stealth assessment predominantly measures a single underlying latent construct, namely, scientific reasoning capability in sound-wave physics, without significant distortion from secondary dimensions or construct-irrelevant noise

Table 2. Standardized Residual Variance Analysis (PCA)

Component	Eigenvalue (Variance)	Empirical (%)	Modeled (%)
Total Raw Variance in Observations	47.50	100.0	100.0
Raw Variance Explained by Measures	17.50	36.8	37.1
Variance Explained by Persons	10.90	23.0	23.2
Variance Explained by Persons	6.60	13.8	13.9
Raw Unexplained Variance (Total)	30.00	63.2	62.9
Unexplained Variance in 1st Contrast	3.03	6.4	-

2. Test Precision and Information Coverage

To further evaluate the targeting and measurement efficiency of the assessment, the Test Information Function (TIF) curve was analyzed (see Figure 3). The TIF curve demonstrates a typical bell-shaped distribution, reaching a maximum information peak of approximately 6.4 at a latent ability level of -0.30 logits. The instrument maintains robust measurement precision (information value > 2.0) across a wide trait range spanning from -2.5 to +2.0 logits. This indicates that the assessment provides high measurement accuracy with minimal standard errors for respondents exhibiting moderate scientific reasoning abilities. Conversely, measurement error increases at the extreme ends of the spectrum (below -4.0 logits and above +4.0 logits), where person targeting becomes less dense.

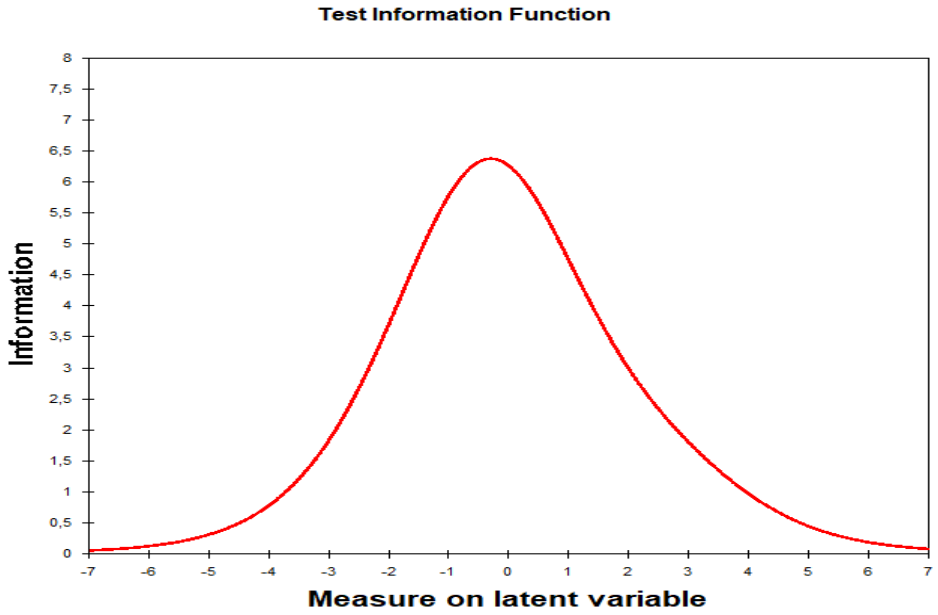


Figure 3. Test Information Function (TIF) curve for the assessment instrument

3. Local Independence

The assumption of local independence dictates that a student's response to a specific item must not influence or dictate their response to any other item after controlling for the primary latent trait. Examination of the largest standardized residual correlations indicated that the maximum observed inter-item residual correlation occurred between item pair Q17 and Q18 (r = 0.47). Because all pairwise residual correlations remained strictly below the critical threshold of r = 0.70, the requirement for local independence was fully satisfied. This demonstrates that each game mission functions as an independent source of evidence within the stealth assessment architecture.

4. Person-Item Targeting

Person-item alignment was evaluated by comparing the mean person ability measure against the mean item difficulty measure (0.00 logit). The mean person measure was calculated at +0.03 logit (SD = 1.37), demonstrating an optimal alignment between student competencies and task difficulty. The spread of item difficulties ranged from -1.95 (Q2; easiest item) to +3.07 (Q22; hardest item), effectively covering the full distribution of student abilities (person measures ranged from -2.85 logit to +3.45 logit).

Internal Consistency and Structural Reliability (RQ2)

To address **RQ2**, summary statistics were calculated to determine the internal consistency, measurement precision, and structural separation capability of the assessment. The Rasch summary statistics is illustrated in Table 3.

Table 3. Rasch Summary Statistics for Persons and Items

Measure Category	Mean (Logit)	SD	Real RMSE
Person Measure	+ 0.03	1.37	0.49
Item Measure	0.00	1.11	0.24

As summarized in Table 3, the instrument demonstrated robust psychometric reliability across both person and item estimations:

1. **Person Reliability and Separation:** The person reliability coefficient was R_p = 0.87 with a person separation index of 2.61. Applying the stratum conversion formula $[H = (4 \times Separation + 1)/3]$ (Bond, 2015; Linacre, 2013) , the instrument successfully stratifies students into at least 3.81 (approximately 4) distinct ability groups (*high achievers, upper-intermediate, lower-intermediate, and low achievers*).
2. **Item Reliability and Separation:** The item reliability coefficient reached an exceptional R_i= 0.95 with an item separation index of 4.60. This high separation value confirms that the sample size (N = 129) was well-suited to establish a highly reproducible item difficulty hierarchy.

Item Diagnostics, Structural Misfits, and Distractor Dynamics (RQ3)

To address **RQ3**, individual item fit statistics were analyzed to detect structural misfits, followed by option-level frequency diagnostics to uncover the underlying causes of erratic response behavior.

1. Identification of Misfitting Items

Evaluating item fit parameters using standard Rasch thresholds (Outfit MNSQ between 0.5 and 1.5, Outfit ZSTD between -2.0 and +2.0, and Point-Measure Correlation r_{pt} > 0.0) revealed that 26 out of the 30 items exhibited excellent structural fit. However, 4 game tasks (Q22, Q25, Q26, and Q30) were flagged as severe statistical outliers (*misfits*) as seen in Table 4.

Table 4. Diagnostic Statistics for Identified Misfitting Items

Entry Number	Total Score	Total Count	JMLE Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEASUR-AL Corr.	Exact Match Obs. (%)	Exact Match Exp. (%)	Item
25	18	129	2.42	0.28	1.51	2.72	5.37	5.24	A-.15	83.7	86.4	Q25
26	26	129	1.88	0.25	1.74	4.64	4.09	5.55	B-.16	70.5	81.8	Q26
30	13	129	2.86	0.31	1.04	0.25	3.72	3.18	C 0.15	89.9	89.9	Q30
22	11	129	3.07	0.34	1.40	1.63	3.16	2.54	D-.10	91.5	91.4	Q22

Entry Number	Total Score	Total Count	JMLE Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEASUR-AL Corr.	Exact Match Obs. (%)	Exact Match Exp. (%)	Item
20	62	129	0.15	0.21	1.46	4.52	1.43	2.61	E 0.25	52.7	72.5	Q20
12	75	129	-0.41	0.21	1.19	1.93	1.11	0.71	F 0.42	67.4	73.2	Q12
8	66	129	-0.02	0.21	1.12	1.32	1.06	0.42	G 0.46	67.4	72.4	Q8
9	64	129	0.07	0.21	0.97	-0.27	1.09	0.68	H 0.52	77.5	72.5	Q9
23	67	129	-0.06	0.21	0.96	-0.45	1.09	0.62	I 0.53	76.0	72.4	Q23
4	84	129	-0.81	0.22	1.02	0.20	1.08	0.47	J 0.49	76.7	75.6	Q4
2	105	129	-1.95	0.26	1.05	0.36	1.04	0.23	K 0.40	82.9	83.7	Q2
7	73	129	-0.32	0.21	1.05	0.53	1.02	0.15	L 0.49	72.1	72.9	Q7
15	79	129	-0.58	0.21	0.98	-0.17	1.05	0.35	M 0.52	75.2	74.1	Q15
19	79	129	-0.58	0.21	1.02	0.24	1.05	0.33	N 0.50	70.5	74.1	Q19
3	79	129	-0.58	0.21	0.96	-0.38	0.97	-0.12	O 0.53	75.2	74.1	Q3
6	61	129	0.20	0.21	0.97	-0.28	0.89	-0.73	o 0.54	72.1	72.5	Q6
16	73	129	-0.32	0.21	0.92	-0.89	0.88	-0.74	n 0.57	76.7	72.9	Q16
28	63	129	0.11	0.21	0.90	-1.19	0.80	-1.44	m 0.59	73.6	72.5	Q28
5	94	129	-1.31	0.23	0.89	-0.97	0.69	-1.37	l 0.57	77.5	79.2	Q5
11	78	129	-0.54	0.21	0.88	-1.32	0.74	-1.72	k 0.60	75.2	73.8	Q11
24	78	129	-0.54	0.21	0.87	-1.43	0.79	-1.33	j 0.60	76.7	73.8	Q24
21	80	129	-0.63	0.21	0.85	-1.58	0.70	-1.96	i 0.62	74.4	74.4	Q21
10	65	129	0.02	0.21	0.83	-2.05	0.76	-1.73	h 0.63	77.5	72.4	Q10
29	81	129	-0.67	0.21	0.79	-2.29	0.67	-2.12	g 0.65	82.9	74.7	Q29
13	76	129	-0.45	0.21	0.78	-2.57	0.67	-2.36	f 0.66	81.4	73.4	Q13
17	67	129	-0.06	0.21	0.78	-2.58	0.68	-2.39	e 0.66	79.1	72.4	Q17
18	60	129	0.24	0.21	0.78	-2.68	0.67	-2.42	d 0.66	78.3	72.6	Q18
1	90	129	-1.10	0.22	0.77	-2.18	0.74	-1.28	c 0.62	86.8	77.6	Q1
27	64	129	0.07	0.21	0.77	-2.80	0.66	-2.57	b 0.67	79.1	72.5	Q27
14	69	129	-0.15	0.21	0.67	-4.16	0.55	-3.63	a 0.73	83.7	72.5	Q14
MEAN	**66.7**	**129.0**	**0.00**	**0.22**	**1.00**	**-0.40**	**1.31**	**-0.16**	—	**76.8**	**75.8**	—
P.SD	**22.0**	**0.0**	**1.12**	**0.03**	**0.24**	**2.03**	**1.15**	**2.18**	—	**7.3**	**5.3**	—

As detailed in Table 4, items Q22, Q25, Q26, and Q30 displayed extremely elevated Outfit Mean Square values (MNSQ > 3.0) and Outfit ZSTD values exceeding +2.0. Crucially, items Q22, Q25, and Q26 exhibited negative Point-Measure Correlations (r_{pt} < 0.0).

2. Option-Level Diagnostic Dynamics (Distractor Analysis)

To investigate the root cause of these misfits, an option-level diagnostic was conducted using the Winsteps *Option Frequencies Table* (Table 5).

Table 5. Option-Level Response Distribution for Misfitting Items Q22 and Q25

Item	Option	Count (N)	Percent (%)	Person Ability Mean (Logit)	Key Status	Diagnostic Finding
Q22	A	12	9%	-0.12	Key (✓)	Endorsed mainly by low-ability students
	B	86	67%	+0.48	Distractor	Plausible Trap / Miskeying suspect
	C	18	14%	-0.35	Distractor	Functioning normally
	D	13	10%	-0.82	Distractor	Functioning normally
Q25	A	15	12%	-0.55	Distractor	Functioning normally
	B	22	17%	-0.22	Key (✓)	Low endorsement by high performers
	C	74	57%	+0.52	Distractor	Strong distractor endorsement by top ability

Item	Option	Count (N)	Percent (%)	Person Ability Mean (Logit)	Key Status	Diagnostic Finding
	D	18	14%	-0.18	Distractor	Functioning normally
Q26	A	18	14%	-0.15	Distractor	Functioning normally
	B	14	11%	+0.42	Distractor	Functioning normally
	C	21	16%	-0.28	Key (✓)	Endorsed mainly by low-ability students
	D	76	59%	+0.45	Distractor	Plausible Trap/Miskeying suspect
Q30	A	32	25%	+0.05	Distractor	Random guessing pattern/poor discrimination
	B	28	22%	+0.02	Distractor	Random guessing pattern/poor discrimination
	C	35	27%	+0.08	Key (✓)	Low endorsement/flat ability distribution
	D	34	26%	-0.01	Distractor	Random guessing pattern/poor discrimination

Discussion

The findings of this study advance the educational assessment literature by demonstrating how stealth assessment can be systematically evaluated using the Rasch Measurement Model to resolve the tension between immersive gameplay and psychometric rigor. Prior research in game-based learning has frequently struggled with the trade-off between player immersion and measurement validity, often relying on classical test theory (CTT) which treats test anxiety and game mechanics as confounding noise (Bonefont et al., 2022; Deng & Liu, 2025). By embedding assessment invisibly within Resonance Odyssey, this study aligns with Flow Theory, confirming that eliminating high-stakes evaluative cues prevents cognitive overload and physics test anxiety without compromising measurement precision.

While previous studies on stealth assessment primarily focused on task completion rates or broad construct validity (Liu et al., 2026; Lu et al., 2025; Shute & Rahimi, 2021), they rarely subjected stealth tasks to rigorous item-level diagnostic analysis under modern Item Response Theory (IRT). The empirical results of this study address this critical research gap. By satisfying the core Rasch assumptions, construct unidimensionality, local independence, and optimal person-item targeting, this research proves that stealth assessment tasks in physics can operate as true invariant measurement units. However, unlike traditional paper-and-pencil physics tests where item misfits are typically attributed to lack of student knowledge, the item-level option diagnosis (Q22, Q25, Q26, and Q30) revealed a unique phenomenon in game-based contexts: high-ability students were frequently tripped up by plausible in-game traps or miskeyed distractors rather than conceptual ignorance. This finding extends current diagnostic assessment frameworks by proving that game mechanics themselves can introduce subtle construct-irrelevant variance if option distractors are not psychometrically calibrated.

The key novelty of this study lies in establishing an empirical bridge between game-based stealth assessment and option-level Rasch diagnostics for physics education. Rather than treating board game actions merely as qualitative learning indicators, this study demonstrates how game-integrated tasks can achieve high item and person reliability while identifying structural flaws in specific task distractors. This provides a replicable methodological roadmap for educators and game designers: stealth assessment in physical or hybrid board games can deliver non-intrusive, anxiety-free measurement while maintaining the highest standard of

psychometric integrity needed for diagnostic evaluations.

CONCLUSION

This study evaluated the psychometric rigor and item-level diagnostics of the 30-item Resonance Odyssey stealth assessment using the Rasch measurement model. The empirical findings confirm that integrating educational measurement directly into board game mechanics successfully satisfies fundamental psychometric criteria: construct unidimensionality (36.8% variance explained), local independence (maximum residual correlation r = 0.47), and high reliability for both items (R_i = 0.95, separation = 4.60) and persons (R_p = 0.87, separation = 2.61). Furthermore, the optimal alignment between student capability and task difficulty (mean person measure of +0.03 logit) demonstrates that non-intrusive stealth assessment can accurately capture scientific reasoning in sound-wave physics while preserving the learning flow and mitigating evaluative anxiety. However, item-level option diagnostics identified four misfitting items (Q22, Q25, Q26, and Q30) with high outfit mean square values (MNSQ > 3.0), driven primarily by unexpected choice endorsements among high-ability students and plausible distractor traps. These results emphasize that establishing overall construct validity is insufficient; granular option-level diagnostics are essential to prevent game-based distractors from introducing construct-irrelevant noise.

Limitations and Future Research

This study is limited by its sample size of 129 participants within a single educational context and a focus on a specific physics topic. Future research should expand the sample size across diverse demographic groups and explore automated, dynamic distractor calibration within physical and hybrid game environments. Overall, Resonance Odyssey provides a viable, psychometrically sound framework for game-based assessment, demonstrating that targeted revision of problematic distractors is key to maximizing diagnostic precision in stealth measurement.

REFERENCES

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Ahmed, A. A. A., Ampry, E. S., Komariah, A., Hassan, I., Thahir, I., Hussein Ali, M., Fawzi Faisal, A., & Zafarani, P. (2022). Investigating the Effect of Using Game-Based Learning on EFL Learners' Motivation and Anxiety. *Education Research International*, *2022*. https://doi.org/10.1155/2022/6503139

Bond, T. (2015). *Applying the Rasch Model* (3rd Editio). Routledge. https://doi.org/https://doi.org/10.4324/9781315814698

Bonefont, V., Carbino, B., & Zakerzadeh, R. (2022). Addressing Test Anxiety Using Game-Based Assessment. *Biomedical Engineering Education*, *2*(2), 319–330. https://doi.org/10.1007/s43683-022-00082-3

Chen, C.-H., Huang, K., & Liu, J.-H. (2020). Inquiry-Enhanced Digital Game-Based Learning: Effects on Secondary Students' Conceptual Understanding in Science, Game Performance, and Behavioral Patterns. *Asia-Pacific Education Researcher*, *29*(4), 319 – 330. https://doi.org/10.1007/s40299-019-00486-w

Deng, Y., & Liu, H. (2025). To overcome test anxiety in on-line assessment:

unpacking the mediator roles of techno competencies, teacher support, self-efficacy, and autonomy. *BMC Psychology*, *13*(1). https://doi.org/10.1186/s40359-025-02545-y

Dumile, N., & Boateng, S. (2026). Game-based learning with Kahoot: Effects on grade 12 learners' anxiety, engagement, and academic performance in genetics. *Eurasia Journal of Mathematics, Science and Technology Education*, *22*(9). https://doi.org/10.29333/ejmste/19124

Fernanda, J. W., & Hidayah, N. (2020). Classical Test Theory dan Rasch Model. *SQUARE : Journal of Mathematics and Mathematics Education*, *2*(1), 49–60.

Linacre, J. M. (2013). *A user's guide to Winsteps Ministeps Rasch-model computer programs*. Winsteps. http://www.winsteps.com/index.htm

Liu, J., Shadiev, R., Ge, D., Cao, M., Xu, Y., & Zhou, R. (2026). Assessing digital etiquette in adolescents: A stealth approach using game-based learning to reduce anxiety, promote flow, and enhance learning gains. *Computers and Education*, *243*. https://doi.org/10.1016/j.compedu.2025.105475

Lu, W., Griffin, J., Sadler, T. D., Laffey, J., & Goggins, S. P. (2025). Game-Based Learning Prediction Model Construction: Toward Validated Stealth Assessment Implementation. *Journal of Learning Analytics*, *12*(1), 293–321. https://doi.org/10.18608/jla.2025.8105

Melander Bowden, H., & Aarsand, P. (2020). Designing and assessing digital games in a classroom: an emerging culture of critique. *Learning, Media and Technology*, *45*(4), 376–394. https://doi.org/10.1080/17439884.2020.1727500

Prieto, M. F., García, C. F., Gómez, M. S., & Sánchez, M. D. P. (2021). Scoring divergent thinking tests: Controversies and solutions . *Revista Iberoamericana de Diagnostico y Evaluacion Psicologica*, *59*(2), 35–47. https://doi.org/10.21865/RIDEP59.2.03

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Skinner, J. (2020). Testing the truth. *Pragmatism Today*, *11*(2), 86–94. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104848194&partnerID=40&md5=e5c5f964dc20740214797b5ed08444cb

Soeharto, S., & Csapó, B. (2022). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, *46*. https://doi.org/https://doi.org/10.1016/j.tsc.2022.101132

Yan, K., Yamada, H., Takagaki, M., & Koizumi, R. (2019). Status ofMathematics Tests in the National Center Test for University Admissions in Japan and How to Improve It. *Juntendo Medical Journal*, *65*(3), 255–260.

You, H. S., Kim, K., Black, K., & Min, K. W. (2018). Assessing science motivation for college students: Validation of the science motivation questionnaire II using the rasch-andrich rating scale model. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(4), 1161–1173. https://doi.org/10.29333/ejmste/81821