

From Play to Validity: A Design–Psychometric–Cognitive Framework for Evaluating Game-based Personality Assessment

Fadillah^{1,3*}, Rahmat Hidayat¹, and Agung Santoso²

¹Faculty of Psychology, Gadjah Mada University, Yogyakarta, Indonesia

²Faculty of Psychology, Sanata Dharma University, Yogyakarta, Indonesia

³Faculty of Art and Design, Institute of Technology Bandung, Bandung, Indonesia

Abstract. Game-Based Assessments (GBAs) offer innovative approaches to measuring psychological constructs by embedding tasks within interactive environments. Despite their potential, concerns remain about validity and whether gameplay reflects stable traits rather than situational or design-driven factors. This study aimed to examine the construct validity of a GBA designed to measure Conscientiousness by integrating cognitive interviewing, Item Response Theory, and design evaluation. Ten participants completed the GBA and participated in follow-up interviews guided by Collins’ framework. The study addressed how gameplay tasks represented Conscientiousness and which design features affected players’ cognitive responses. Findings showed that early stages (1–2) demonstrated strong alignment, clear instructions, understandable goals, and items with adequate fit and discrimination, eliciting structured, trait-consistent behaviors. Later stages (3–6) exhibited weaker psychometric performance, divided attention, and confusion from visual similarity. Overall, results highlight the need to align psychometric rigor, cognitive clarity, and design simplicity in developing valid GBAs.

1 Introduction

Game-Based Assessment (GBA) embeds psychometric constructs into game environments, combining aesthetics, narrative, and mechanics to elicit behavioral expressions of traits [1]. Unlike self-reports, GBAs are both psychometric tools and design experiences, making validation dependent on measurement and design factors. This study examines a GBA developed to assess Conscientiousness through a theory-driven, Design–Play–Experience framework [2]. Earlier testing showed convergent validity with IPIP-NEO-120 ($\rho = .350$, $p < .001$), with strong discrimination in Stages 1–2 but weak or misfitting items in later stages, and modest reliability. To explore why performance varied across stages, we applied cognitive interviewing based on Collins’ framework [3], assessing comprehension, judgment, response, layout, and authenticity. Integrating IRT results with qualitative insights

* Corresponding author: fadillah@mail.ugm.ac.id

provided a dual perspective, highlighting where the design supported construct representation and where refinement is needed.

2 Methods

2.1 Participants

The study included ten participants (5 females, 5 males) aged 21–30 years, with educational backgrounds spanning undergraduate (Psychology, Design, Business, Communication) and graduate studies (Computer Science, Education, Engineering, Information Systems). Gaming experience varied from casual mobile and puzzle games to moderate engagement with simulation and mixed genres, as well as frequent role-playing, strategy, or competitive play.

2.2 Instruments

2.2.1 Game-Based Assessment of Conscientiousness

The instrument under investigation was a GBA measuring Conscientiousness, chosen because it is widely recognized as the strongest predictor of job performance among the Big Five traits. Developed through an iterative, design-driven process, the GBA operationalized Conscientiousness facets, such as orderliness, dutifulness, and achievement-striving, into interactive gameplay scenarios [2]. Previous psychometric analyses provided evidence of convergent validity with the IPIP-NEO-120 Conscientiousness scale and examined internal structure using Item Response Theory (see Table 1), revealing strong discrimination in early stages but weaker or misfitting in later stages.

Table 1. GBA IRT Analysis

	a	Note	b1	b2	b3	Note
Stage1	4,203577	High	-2,50905	-1,6228	-0,2184216	Well-ordered
Stage2	3,423423	High	-1,34588	-0,7198	1,2412667	Well-ordered
Stage3	0,22149	Low	-10,2379	2,013364	6,2536058	Extreme
Stage4	0,031483	Very Low	-30,7977	-18,2659	45,3258686	Extreme
Stage5	-0,04594	Misfitting	47,4814	-5,28204	NA	Nonesteemable
Stage6	0,211035	Low	-16,1388	-2,08394	9,4540936	Well-ordered

2.2.2 Cognitive Interviewing Protocol

Qualitative data were collected using a structured cognitive interview guide adapted from Collins’ framework [3], covering six domains. The guide was tailored to the GBA and organized by stage of gameplay. These probes ensured that each cognitive domain was systematically addressed while allowing participants to elaborate freely.

2.3 Data Analysis

Thematic analysis was conducted using Collins’ six-domain cognitive interviewing framework, with iterative, inductive coding to capture both common themes and unique

perspectives. Verbatim quotations were included to illustrate findings, and qualitative insights were triangulated with psychometric results.

3 Results

3.1 Stage 1

At Stage 1, participants encountered the initial gameplay. The analysis focuses on how clearly they understood the game’s purpose and how they retrieved and applied information.

Table 2. Cognitive Interview Result (Stage 1)

Domain	Evidence (Verbatim Quotes)	Interpretation
Comprehension	“The instructions at the beginning of the game were clear and easy to understand” (S4); “Quite clear and easy to follow” (S8)	Initial comprehension worked well; most participants quickly understood the goal.
Retrieval	“No need for additional tutorials” (S1); “Clear, but could be improved with red dots as guidance” (S7)	Basic information was sufficient, though some suggested extra visual symbols for clarity.
Judgment	“There was nothing confusing, just follow the instructions” (S3); “At first I went left and hit a dead end, then realized I had to go right” (S6)	Decisions were generally smooth, with minor trial-and-error in navigation.
Response	“I chose the easiest option” (S1); “I picked the shortest path to save effort” (S3)	Responses were practical and simple at the start.
Burden & Layout	“Too many instructions when buying items, maybe use a button” (S7)	Layout was helpful, but excessive pop-ups added unnecessary load.
Authenticity & Engagement	“The initial instructions were enough to help” (S6)	Early gameplay supported focus, though narrative visuals could be richer for engagement.

In the initial stage (Table 2), participants generally understood the game flow and found instructions clear (S4, S8). Decision-making was straightforward (S3), though minor trial-and-error occurred (S6), with many favoring simple choices (S1). One suggested visual markers for guidance (S7), while excessive instructions during equipment purchasing felt burdensome. Overall, the stage ensured clarity but would benefit from streamlined instructions and stronger visual cues.

3.2 Stage 2

In Stage 2, players began making basic choices. Here, the interview analysis highlights the clarity of decision points, and whether the design encouraged meaningful engagement (Table 3).

Table 3. Cognitive Interview Result (Stage 2)

Domain	Evidence (Verbatim Quotes)	Interpretation
Comprehension	“I chose the easiest option” (S1); “I went straight to the result without overthinking” (S2)	Comprehension of tasks was strong, with a tendency to favor simple options.
Retrieval	“The information was clear to help with decisions” (S4); “Very helpful, it explained the consequences” (S6)	On-screen info was actively used to anticipate outcomes.

Judgment	“No icons or symbols made me doubt” (S6); “In preparation, shouldn’t food be included?” (S7)	Most judgments were clear, though context details required clarification.
Response	“I preferred safe and normal decisions” (S5)	Responses leaned toward risk avoidance.
Burden & Layout	“The layout was clear enough” (S6); “The equipment shop needs clearer guidance” (S7)	Layout supported comprehension, though item preparation remained less intuitive.
Authenticity & Engagement	“I could choose whatever items I wanted to bring” (S6)	The stage offered room for personal choice, boosting engagement.

In Stage 2, participants showed good comprehension and favored straightforward, practical choices (S1, S2, S5). On-screen information was clear and non-confusing (S4, S6), though one noted missing scenario details (S7). The freedom to select items was engaging (S6), indicating the stage effectively supported simple, construct-relevant decision-making.

3.3 Stage 3

The analysis shows how participants relied on different strategies and the extent to which gameplay reflected stable personality preferences (Table 4).

Table 4. Cognitive Interview Result (Stage 3)

Domain	Evidence (Verbatim Quotes)	Interpretation
Comprehension	“Clear enough, not too complicated” (S6)	Comprehension was still intact despite growing complexity.
Retrieval	“I planned the order first” (S4); “I just placed items as they appeared” (S1); “I adjusted to the visible pattern” (S6)	Strategies varied (systematic vs. spontaneous), often based on improvisation.
Judgment	“It was difficult when my focus was divided” (S8)	Difficulty arose from divided attention, showing weak discrimination of stable preferences.
Response	“Sometimes I rushed and made mistakes” (S5)	Responses were situational and error-prone, less tied to personality tendencies.
Burden & Layout	“It was difficult when my focus was divided” (S8)	Cognitive load increased mainly due to visual complexity.
Authenticity & Engagement	“I tried to adapt to the pattern I saw” (S6)	Strategies reflected short-term adaptation rather than stable personality preferences.

As complexity increased, participants employed varied strategies, ranging from pre-planning (S4) to spontaneous placement (S1) or adapting to emerging patterns (S6). Several noted difficulties when attention was divided (S8) or when rushing led to errors (S5). These patterns indicate that responses were shaped more by situational demands than by stable personality tendencies, reducing discriminant power at higher complexity levels.

3.4 Stage 4

At Stage 4 (Table 5), decision-making tasks became more prominent. The analysis illustrates how participants processed options, their confidence levels, and how visual similarity influenced responses.

Table 5. Cognitive Interview Result (Stage 4)

Domain	Evidence (Verbatim Quotes)	Interpretation
Comprehension	“I followed the instructions and didn’t deviate” (S7)	Rules were understood consistently.
Retrieval	“I chose according to a logical sequence” (S3); “Sometimes I guessed first and corrected later” (S1)	Decision-making styles diverged: logical vs. trial-and-error.
Judgment	“Confident because I followed my sequence” (S4); “Sometimes unsure, especially when options looked similar” (S6)	Confidence varied; doubts were triggered more by visual similarity than personal preference.
Response	“Sometimes I guessed first, then corrected if wrong” (S1)	Responses were improvisational, more design-driven than personality-driven.
Burden & Layout	“Sometimes unsure, especially when options looked similar” (S6)	Visually similar options increased cognitive load, limiting discrimination of preferences.
Authenticity & Engagement	“I followed the instructions and didn’t deviate” (S7)	Context felt realistic, but outcomes leaned more on visuals than personality cues.

The decision-making stage showed contrasting response styles: some participants followed a logical sequence (e.g., S3), while others relied on improvisation (e.g., S1). Confidence also varied, from assured choices based on strategy (S4) to uncertainty when options appeared visually similar (S6). Visual similarity increased cognitive load and limited this stage’s capacity to reflect personality preferences, as decisions were shaped more by design features than by stable individual tendencies.

3.5 Stage 5

Stage 5 produced more polarized feedback. Some participants enjoyed the challenge, while others felt overwhelmed by visual clutter.

Table 6. Cognitive Interview Result (Stage 5)

Domain	Evidence (Verbatim Quotes)	Interpretation
Comprehension	“It was okay, not too difficult” (S6)	Comprehension remained, though gameplay became more complex.
Retrieval	“I just placed things as they appeared” (S1)	Intuitive strategies dominated, shaped by visuals rather than deliberation.
Judgment	“Fun and challenging” (S2); “Sometimes too crowded, so confusing” (S8)	Mixed judgments; responses were based more on aesthetics/visual clutter than preferences.
Response	“Sometimes it got confusing” (S8)	Responses were inconsistent, driven by visual overload.
Burden & Layout	“Colors and icons helped” (S7); “Sometimes too crowded” (S8)	Visuals supported some but overwhelmed others, leading to reversed psychometric scores.
Authenticity & Engagement	“Fun and challenging” (S2)	Engagement arose, but mostly situational and not tied to stable personality expression.

At this stage (Table 6), participants expressed divided experiences: some found the gameplay enjoyable and challenging (S2), while others reported confusion due to visual overload (S8). Although certain visual elements such as colors and icons were supportive (S7), they also increased distraction. These findings indicate that visual complexity, rather

than personality traits, primarily shaped responses, consistent with the reversed psychometric patterns observed.

3.6 Stage 6

Stage 6 was reflective. Participants were asked to consider whether the game mirrored their real-world behaviors. Responses were mixed, suggesting that representational validity varied considerably.

Table 7. Cognitive Interview Result (Stage 6)

Domain	Evidence (Verbatim Quotes)	Interpretation
Comprehension	“Quite similar to how I usually organize tasks” (S4); “Not really, I’m usually more relaxed” (S1)	Perceived relevance differed widely, indicating inconsistent alignment.
Retrieval	“Yes, it reflected how I usually manage” (S6)	Some felt it matched their real habits, others did not.
Judgment	“Not really, I’m usually more relaxed” (S1)	Evaluations were split; preference representation remained inconsistent.
Response	“The visual instructions should be clearer” (S8)	Feedback highlighted design clarity as a barrier.
Burden & Layout	“The visual instructions should be clearer” (S8)	Visual guidance was seen as insufficient, affecting perceived validity.
Authenticity & Engagement	“Yes, it reflected how I usually manage” (S6)	The game represented real behavior for some, but not universally.

In the reflection stage (Table 7), participants expressed mixed views on the game’s relevance to their personal habits. Some perceived strong alignment with their task management style (e.g., S4, S6), while others highlighted design issues such as unclear visual instructions (e.g., S1, S8). Overall, reflections revealed variability in perceived authenticity, indicating that the game did not consistently capture personality-related behaviors.

Table 8. Psychometric-Cognitive-Design Analysis

Stage	Psychometric (IRT)	Cognitive (Interview)	Design (Gameplay/Visual)
1	Adequate fit & moderate discrimination	Clear comprehension, safe/simple decisions	Simple layout, clear feedback
2	Adequate fit & moderate discrimination	Clear decisions, risk-averse choices	Clear info, item choice engaging
3	Low discrimination, weaker thresholds	Situational, divided focus, rushed errors	Visual similarity increases load
4	Low discrimination, weak thresholds	Mixed strategies, doubts with similar options	Similar options drive hesitation
5	Poor fit, reversal effects	Split opinions, confusion due to clutter	Visual overload dominates responses
6	Low discrimination, weak thresholds	Mixed reflections, inconsistent authenticity	Visual clarity issues, limited realism

Taken together, the results indicate a clear contrast between the early and later stages of the GBA (see Table 8). Stages 1–2 facilitated comprehension, decision-making, and engagement, aligning with the intended measurement goals, whereas Stages 3–6 were increasingly influenced by situational complexity and visual design, reducing discriminant power. Notably, Stage 5 showed a reversal effect, with responses driven more by visual overload than trait-relevant behavior.

4 Discussion

The evaluation combining cognitive interviews and IRT analysis provides critical insights into the instrument's capacity to measure Conscientiousness. Interpreting the findings through psychometric, cognitive, and design perspectives [1], [4], [5] clarifies both strengths and areas requiring refinement, consistent with recent GBA research [6].

Qualitative and quantitative evidence showed that Stages 1–2 effectively elicited construct-relevant responses. Participants reported clear instructions and straightforward decisions, reflecting stable, structured tendencies [4], while IRT analyses indicated adequate fit and moderate discrimination [2]. These findings align with prior research demonstrating that early, low-complexity tasks in GBAs foster engagement and yield stronger trait-related behavior [5] [7]. From a design perspective, simple layouts and clear feedback reduced extraneous variance, consistent with guidelines emphasizing clarity and low cognitive load [1], [8], and meta-analytic evidence supports stronger validity for theory-driven, low-complexity designs [9]. As task complexity increased in Stages 3–4, participants relied more on situational strategies, with some planning systematically and others responding spontaneously, leading to more errors and confusion. IRT results confirmed lower discrimination and weaker thresholds, consistent with evidence that complexity introduces situational noise [10]. Design factors such as visual similarity further heightened ambiguity and cognitive load, reducing measurement fidelity [11], a risk also highlighted in the Game-Based Assessment Framework [12].

Stage 5 posed the greatest challenges, with participants split between finding the gameplay enjoyable and reporting overload from excessive visuals. While colors and icons were sometimes helpful, they often produced clutter that distracted from trait-relevant decision-making. Psychometric evidence confirmed this stage's weakness, showing poor fit, low discrimination, and reversal effects [2]. These results align with research showing that overemphasis on gamification and visual cues can reduce construct validity and increase susceptibility to distortion [13]. From a design perspective, Stage 5 demonstrates how non-construct variance, such as tolerance for visual clutter, can overshadow trait-driven responses [14]. Stage 6 prompted reflection on real-world relevance, producing mixed responses: some saw alignment with their task management, while others disagreed or cited unclear visuals. IRT results showed unstable discrimination, consistent with critiques that reflective or narrative-heavy stages may lack fidelity [15]. Recent GBA applications also stress the importance of participant feedback on authenticity for representational validity [6].

Across psychometric, cognitive, and design lenses, findings show the GBA performs well in early stages but weakens later. Validity is strongest when discrimination, comprehension, and design clarity align, as in Stages 1–2, but deteriorates in Stages 3–6 due to situational and design-driven noise [11], [12], [15]. This aligns with broader evidence that GBA quality depends on task design, cognitive load, and feedback [5], [8]. Meta-analytic and experimental studies further confirm that simpler, constrained GBAs yield stronger construct correlations and reduce susceptibility to distortion [9], [13].

5 Conclusion

The evaluation of the GBA through cognitive interviews, IRT, and design analysis revealed strong alignment in early stages but reduced precision in later ones, with Stage 5 showing reversal effects due to visual overload. While the GBA demonstrates promise for assessing Conscientiousness, limitations include a small, non-representative sample and design-specific psychometric outcomes. Future work should refine design features to balance engagement with construct validity and leverage multimodal data to enhance measurement precision.

References

1. Y. J. Kim and D. Ifenthaler, "Game-Based Assessment: The Past Ten Years and Moving Forward," 2019, pp. 3–11. doi: 10.1007/978-3-030-15569-8_1.
2. Fadillah, R. Hidayat, and A. Santoso, "Designing game-based assessment for conscientiousness: A design–play–experience approach," in *Proceedings of the International Conference on Serious Games and Edutainment*, 2024.
3. D. Collins, *Cognitive Interviewing Practice*. 2015. Doi:10.4135/9781473910102
4. F. Luthans and C. M. Youssef-Morgan, "Psychological Capital: An Evidence-Based Positive Approach," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 4, no. 1, pp. 339–366, Mar. 2017, doi: 10.1146/annurev-orgpsych-032516-113324.
5. N. Gazit, G. Ben-Gal, and R. Eliashar, "Game-Based Assessment of Cognitive Abilities and Personality Characteristics for Surgical Resident Selection: A Preliminary Validation Study," *JMIR Med Educ*, vol. 11, pp. e72264–e72264, Aug. 2025, doi: 10.2196/72264.
6. F. Y. Wu, E. Mulfinger, L. Alexander, A. L. Sinclair, R. A. McCloy, and F. L. Oswald, "Individual differences at play: An investigation into measuring Big Five personality facets with game-based assessments," *International Journal of Selection and Assessment*, vol. 30, no. 1, pp. 62–81, Mar. 2022, doi: 10.1111/ijsa.12360.
7. R. Landers, "Game-based assessments: Design and validation [Keynote]," *Game-based assessment: An interdisciplinary workshop*, University of Minnesota, Aug. 2019.
8. T. Bipp, S. Wee, M. Walczok, and L. Hansal, "The Relationship Between Game-Related Assessment and Traditional Measures of Cognitive Ability—A Meta-Analysis," *J Intell*, vol. 12, no. 12, p. 129, Dec. 2024, doi: 10.3390/jintelligence12120129.
9. F. Lievens and P. R. Sackett, "The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance.," *Journal of Applied Psychology*, vol. 97, no. 2, pp. 460–468, 2012, doi: 10.1037/a0025741.
10. R. S. J. d. Baker, S. K. D’Mello, Ma. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments," *Int J Hum Comput Stud*, vol. 68, no. 4, pp. 223–241, Apr. 2010, doi: 10.1016/j.ijhcs.2009.12.003.
11. C. Udeozor, F. R. Abegão, and J. Glassey, "Measuring learning in digital games: Applying a game-based assessment framework," *British Journal of Educational Technology*, vol. 55, no. 3, pp. 957–991, May 2024, doi: 10.1111/bjet.13407.
12. M. L. Ohlms, K. G. Melchers, U. P. Kanning, and A. J. Barends, "Game on, Faking off? Are Game-Based Assessments Less Susceptible to Faking Than Traditional Assessments?," *J Bus Psychol*, Apr. 2025, doi: 10.1007/s10869-025-10019-6.
13. W. J. Arthur, W. J. Bennett, and A. S. Huffcutt, "Visual and multimedia-based situational judgment tests: Theory, design, and psychometric issues," *Hum Perform*, vol. 27, no. 5, pp. 405–431, 2014.
14. D. Balsis, T. Woods, and K. Gleason, "Validity and fidelity in multimedia-based assessments of personality," *Personality Assessment Quarterly*, vol. 34, no. 2, pp. 115–130, 2019.
15. G. Serio, M. Balsamo, and L. Carlucci, "Game-Based Assessment: Between Goals and Psychometric Rigor," 2025, pp. 192–204. doi: 10.1007/978-3-031-81706-9_14.