International Journal of Serious Games

ISSN: 2384-8766
https://journal.seriousgamessociety.org/

Article

Convergent Validity of Game-Based Assessment: A Meta-Analysis

Fadillah[1,3], Rahmat Hidayat[1] and Agung Santoso[2]

[1]Faculty of Psychology, Gadjah Mada University, Yogyakarta, Indonesia; [2]Faculty of Psychology, Sanata Dharma University, Yogyakarta, Indonesia, [3]Faculty of Art and Design, Bandung Institute of Technology, Bandung, Indonesia
fadillah@itb.ac.id

Keywords:

Serious game
Game-based assessment
Meta-analysis
Validity
Assessment tools
Psychometric evaluation

Received: March 2025
Accepted: August 2025
Published: October 2025
DOI: 10.17083/etxn7795

Abstract

Game-Based Assessments (GBAs) have emerged as innovative tools for measuring personality traits, particularly in recruitment and employee selection. This meta-analysis aims to evaluate the convergent validity of GBAs compared to traditional self-report personality measures, addressing ongoing concerns about their psychometric robustness. A total of 18 studies from 13 peer-reviewed articles (2002–2025) were systematically reviewed using data from Scopus, ProQuest, Wiley Online Library, and ScienceDirect. Random-effects modeling, heterogeneity analysis, and publication bias tests were conducted, with sample size, game type, and statistical method examined as potential moderators. Findings revealed a moderate and statistically significant overall effect size ($r = .516$, $Z = 8.088$, $p < .001$), indicating meaningful convergence between GBAs and self-report measures. Despite this, substantial heterogeneity across studies was observed, with no significant moderation effects and minimal evidence of publication bias. This study offers the first comprehensive meta-analytic synthesis of GBA convergent validity, contributing original empirical support for their utility while critically highlighting conceptual issues such as circular validation and the absence of standardized frameworks. The impact of this research lies in its advancement of the psychometric foundations for GBAs and its call for future validation efforts using methods like item response theory and predictive designs linked to behavioral outcomes.

1. Introduction

Game-Based Assessment (GBA) has emerged as an innovative alternative to traditional self-report questionnaires, leveraging the principles of game design to evaluate human characteristics such as skills, knowledge, and personality. Unlike conventional self-report measures, which rely heavily on individuals' self-perceptions and are prone to biases such as social desirability and faking, GBA involves performance-based assessments. Through

gameplay, participants engage in tasks and challenges that elicit spontaneous behaviors, providing insights into personality traits in real-time and within context. This approach is less dependent on participants' ability to recall or perceive themselves accurately and offers a dynamic way to assess traits while reducing the influence of faking responses [1], [2].

Implicit Trait Policies (ITP) theory explains the link between player behavior and personality in GBA as trait-driven beliefs about the effectiveness of responses in specific situations [3]. When individuals are presented with choices, their selected responses often align with their underlying personality characteristics [4]. In GBAs, game elements present scenarios and choices designed to elicit authentic, trait-driven responses, enhancing their utility for personality assessment [5].

The immersive and interactive nature of GBA not only enhances the ecological validity of assessments but also fosters motivation and engagement, encouraging participants to enter a state of flow and reducing test anxiety [6]. These gameful features make GBA particularly appealing in high-stakes contexts, such as pre-employment assessments and educational settings, where engagement and accurate data collection are critical [7]. The growing adoption of GBA in Industrial-Organizational Psychology (IOP) research highlights its potential benefits in assessing psychological attributes in work-related contexts, compared to traditional methods, which are often lengthy and stress-inducing [8]. To design effective GBA that elicit authentic and personality-driven responses, several important factors that should be considered to make sure the assessments are both engaging and valid. A factor that is frequently discussed in previous research is aligning the game design with the assessment goals. This involves developing game elements that accurately measure the traits being assessed while maintaining the integrity of the assessment process [9], [10]. In other words, the approach to designing the game and its structure plays a significant role. Additionally, it is also important to manage cognitive load and complexity, as well as to incorporate psychometric principles to ensure the assessment's reliability and validity [5], [11], [12].

1.1 Theory-driven and Data-driven design approaches

One of critical factor influencing the effectiveness of GBA is its design approach. At least two main strategies have been employed in the creation of GBAs, theory-driven and data-driven approach [9]. However, the choice of approach impacts the balance between conceptual clarity and practical adaptability, raising important questions about the robustness and generalizability of GBA tools. A theory-driven approach is grounded in established psychological frameworks and theories, ensuring that the game scenarios, tasks, and response mechanisms are systematically aligned with the constructs being measured. For example, if a GBA is intended to assess traits like conscientiousness or extraversion, a theory-driven approach would design tasks that explicitly reflect behaviors associated with these traits, guided by psychological models such as the Big Five personality framework. This alignment helps ensure construct validity by directly linking game-based behaviors to theoretical underpinnings. For instance, Landers and Collmus developed a narrative-driven GBA by converting a traditional personality measure into an interactive story-based format, explicitly targeting Big Five traits through contextually embedded decision points [13]. Similarly, Barends et al. designed an assessment game grounded in the HEXACO model to measure Honesty-Humility, ensuring alignment between player behaviors and the intended psychological construct [14]. On the other hand, a data-driven approach relies on empirical data to design and refine the assessment. For example, Ramos-Villagrasa and Fernández-Del-Río developed a gameful assessment whose predictive validity was evaluated through applicant reactions and real-world outcomes [15]. Likewise, Wu et al. employed data analytics to explore how micro-behaviors during gameplay could reflect specific Big Five personality facets [16]. By collecting and analyzing large datasets of gameplay behavior, patterns that correlate with specific psychological traits can emerge, allowing the assessment to adapt and optimize based on real-world evidence. While this

approach can enhance predictive validity and uncover nuanced relationships not captured by existing theories, it risks lacking a clear conceptual basis. Without theoretical guidance, there is a potential for misinterpreting gameplay behaviors or overfitting the assessment to specific datasets, which may limit its generalizability.

1.2 Game-Based Assessment and Gamified Assessment

In the context of game-related assessments, Game-Based Assessment and Gamified Assessment represent two distinct approaches, each influencing the convergent validity index in different ways [9]. First, Game-Based Assessment refers to an assessment method in which job candidates engage in a structured gameplay experience. They participate in a core gameplay loop where their behaviors and decisions provide insight into their psychological traits. In GBA, the tasks, scenarios, and choices presented within the game are integral to the assessment process itself, meaning that the gameplay serves as a direct measure of psychological traits such as personality. The behaviors participants exhibit while navigating the game provide authentic, real-time data, allowing for more accurate and context-rich assessments of their personality or other attributes [2].

Second, gameful design refers to practices by assessment professionals who use game mechanics or other game concepts to guide decision-making during the assessment design process. In this case, game elements are incorporated not necessarily to create a new form of assessment, but to structure existing assessments with game-inspired frameworks, which may enhance engagement or participant motivation. However, the assessment still primarily follows traditional formats and methods, with game mechanics serving as tools to influence the assessment experience, not as core measurement devices. Third, gamification refers to the practice of applying game mechanics or concepts to redesign existing assessments, such as questionnaires or surveys. Both gameful design and gamification are gamified assessments that based on traditional methods but incorporate elements like points, levels, badges, or rewards to make the process more engaging and motivating. Although this can enhance participants' involvement and reduce test anxiety, it does not fundamentally change the nature of the assessment itself, which still relies heavily on self-reported data or other traditional forms of measurement [17].

While Game-Based Assessment is itself a core assessment method, gameful design and gamification are more accurately described as redesign strategies that incorporate game-inspired elements into existing structures. These approaches impact the convergent validity index differently. In the case of GBA, since the game is designed to measure specific psychological traits through in-game behaviors, the convergent validity index may be higher because the data collected from GBA is more likely to reflect authentic, behavioral responses that align directly with established personality measures, that might be enhancing construct validity.

In contrast, gamified assessments and gameful designs typically have a weaker direct connection to the psychological traits being measured. While these approaches can improve engagement and motivation, they do not fundamentally change the measurement method. Instead, they often rely on self-reporting or traditional metrics, which may not capture behavioral data as accurately as GBA. This difference suggests that gamification enhances the assessment experience but does not necessarily improve the validity of the psychological constructs being measured.

In summary, the primary distinction between Game-Based Assessment and Gamified Assessments lies in the depth of integration of game elements within the assessment process. GBA serves as a direct method of measurement, while gamified assessments are strategies to enhance existing assessments. This distinction significantly impacts the convergent validity

index. There is a need for further research to explore how these approaches can be optimized to ensure that they accurately capture psychological traits. This gap in the literature underscores the importance of conducting studies that investigate the effectiveness of these assessment types, particularly in terms of their validity, to inform best practices and guide future developments in the field of psychological measurement.

1.3 Multiple personality and Single personality traits assessment

In particular, GBA has been used to assess multiple personality traits, such as those in the Big Five personality model i.e extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness. However, a key issue arises in determining how many personality attributes can effectively be assessed within a single game, and how this impacts the validity of the results. While assessing multiple personality traits in one game might seem advantageous, it introduces potential challenges related to the game's complexity and the overlap of traits being measured [18].The more attributes a game attempts to assess, the more intricate the design must be to ensure each trait is adequately measured without muddling or overshadowing other traits. For instance, if a game assesses extraversion, conscientiousness, and openness at once, the behaviors triggered by the game may be influenced by more than one trait simultaneously, making it difficult to isolate the specific influence of each. Furthermore, as the number of traits increases, so does the risk of creating cognitive overload for the player [16]. This presents a significant gap in current research, as it remains unclear how the number of personality attributes assessed in a single game influences the convergent validity of GBA scores.

As the design of GBA becomes more complex with the inclusion of multiple traits, the potential for overlapping, misaligned, or inaccurate measures increases, which could weaken the convergent validity of the scores. Despite the growing body of research into GBA, studies that explore the relationship between the number of personality traits assessed and the validity of the results remain limited. This gap in understanding necessitates a more systematic investigation into the factors that influence the validity of GBA, particularly as it pertains to the number of personality attributes measured.

1.4 Pearson Correlation and Regression Analysis

To determine the validity of Game-Based Assessment (GBA), researchers often use statistical methods such as Pearson correlation and regression analysis. These techniques help assess how GBA scores relate to established personality assessments or relevant outcomes, providing insight into its effectiveness as a measurement tool [9]. While both methods are valuable, they vary in complexity and the type of insights they offer.

Pearson correlation is widely used in GBA research due to its simplicity and ease of understanding. It measures the strength and direction of the linear relationship between two variables, such as GBA scores and scores from traditional assessments like self-report questionnaires. Researchers use Pearson correlation to examine convergent validity, as it provides a straightforward metric of how closely GBA aligns with existing, validated measures of the same traits [19]. Because Pearson correlation is easy to calculate and interpret, it is often the go-to choice for many professionals in the field. Its results are intuitive, with higher correlation coefficients indicating stronger relationships and thus stronger evidence of validity. This simplicity makes Pearson correlation accessible to a wide range of researchers and practitioners, even those without extensive statistical training. In contrast, regression analysis offers a deeper and more complex method for analyzing the validity of GBA. While Pearson correlation provides a measure of association, regression goes further by exploring how well GBA scores can predict other relevant outcomes, such as job performance or behavioral traits

[20], [21] . Multiple regression can include GBA scores as one of several predictors and assess the unique contribution of GBA in explaining variance in a dependent variable. For example, regression can determine how well GBA scores predict real-world behaviors or outcomes in addition to traditional personality measures. This ability makes regression a powerful tool, as it provides insights not only into relationships but also into the predictive power of GBA [22]. However, regression analysis is more complicated to perform and interpret than Pearson correlation. It requires careful consideration of model assumptions, potential confounding variables, and the complexity of interactions between multiple predictors. As a result, regression analysis may require more sophisticated statistical training and expertise to apply correctly and meaningfully [23].

Despite the valuable insights that both Pearson correlation and regression analysis provide, there remains a significant gap in the literature regarding the comprehensive analysis of GBA validity. Most studies on GBA validity have primarily relied on Pearson correlation to assess convergent validity, focusing on simple associations between GBA and traditional personality measures [9], [24]. Moreover, while regression analysis offers a more sophisticated approach to understanding predictive validity, its complexity has led to fewer studies utilizing it in GBA research. This complexity often deters researchers from employing regression as a primary method, leaving a gap in understanding how GBA can predict meaningful outcomes beyond simple correlations.

However, to ensure the validity reliability of GBA as a personality assessment tool, a comprehensive review of existing studies is necessary. A meta-analysis is particularly valuable in this context, as it can aggregate findings from various studies to provide a clearer, more reliable understanding of GBA's convergent validity [25]. Given the relatively limited number of studies on GBA's validity, conducting a meta-analysis will help identify patterns and establish a more generalized assessment of its effectiveness. By synthesizing the existing body of work, this approach will provide insights into whether GBA truly measures personality traits in the same way as traditional methods, helping to refine its design and application in the future [19].

Therefore, this study aims to systematically evaluate the convergent validity of GBA through a meta-analysis of existing literature. To what extent do GBA demonstrate adequate convergent validity? By synthesizing prior research, it will examine the strength and consistency of relationships between GBA scores and traditional personality assessments. Additionally, this study seeks to identify potential moderators, such as differences in game design or assessment contexts, that may influence GBA's effectiveness. The findings will contribute to enhancing validity of GBA, offering valuable insights for both researchers and practitioners in the field of psychological assessment.

2. Methods and Material

2.1 Sample Criteria

This study employed a meta-analytic approach in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [26]. Meta-analysis was selected as the methodological framework due to its capacity to synthesize findings across multiple empirical studies and provide a more comprehensive understanding of the convergent validity of GBAs. The use of PRISMA guidelines ensured transparency, methodological rigor, and reproducibility throughout the study identification, screening, and inclusion processes. This meta-analysis process was publicly accessible on OSF at https://doi.org/10.17605/OSF.IO/B6WJM. The criteria used for sample articles in this study were:

1. Articles were written in English with full-text access.
2. The participants in the studies included should be adults in pre-employment settings who are not undergoing clinical therapies or interventions.
3. The psychometric evaluation only employed on GBA measurements and not included GBL (Game-based Learning) which has different objectives. The primary emphasis in GBA should be directed solely towards assessment, with no emphasis on discovering learning methods. While learning may naturally occur as a positive outcome of GBA, it is not the primary objective. The design of gameplay elements aligns with these respective objectives.
4. Only application of GBA on personality assessment that would be included. Articles included application of GBA on any other psychological assessment, such as cognitive evaluation or skills, were excluded. Measurement of personality aspects is limited to aspects of character or personality which are carried out through psychometric studies with standardized scales or questionnaires. Concepts outside the psychological aspects such as cognitive and skills will be excluded.
5. The study design was constrained to quantitative studies, which included experimental or correlation studies. Qualitative study designs included analysis report resulted from qualitative data, reviews or theoretical studies wouldbe excluded.
6. Statistical information showed correlation coefficients between GBAs and scale scores derived from validated self-report instruments.

2.2 Search Strategy

To carry out the literature search in the present study, the authors employed Scopus, ProQuest, Willey Online Library, and ScienceDirect to identify publication journals published between 2002 – January 2025 as refers to Figure 1. Terms or sets of keywords were selected in accordance with the research question. To streamline the search process, Boolean operators were employed: ("game-based assessment" OR "game assessment" OR "gaming" OR "serious game") AND ("psychological" OR "personality trait" OR "personality" OR "characteristic" OR "character" OR "trait") NOT (cognitive) NOT (skills) AND ("validity" OR "reliability" OR "effectiveness" OR "accuracy" OR "trustworthiness") AND ("evaluation" OR "measurement" OR "testing" OR "assessment").

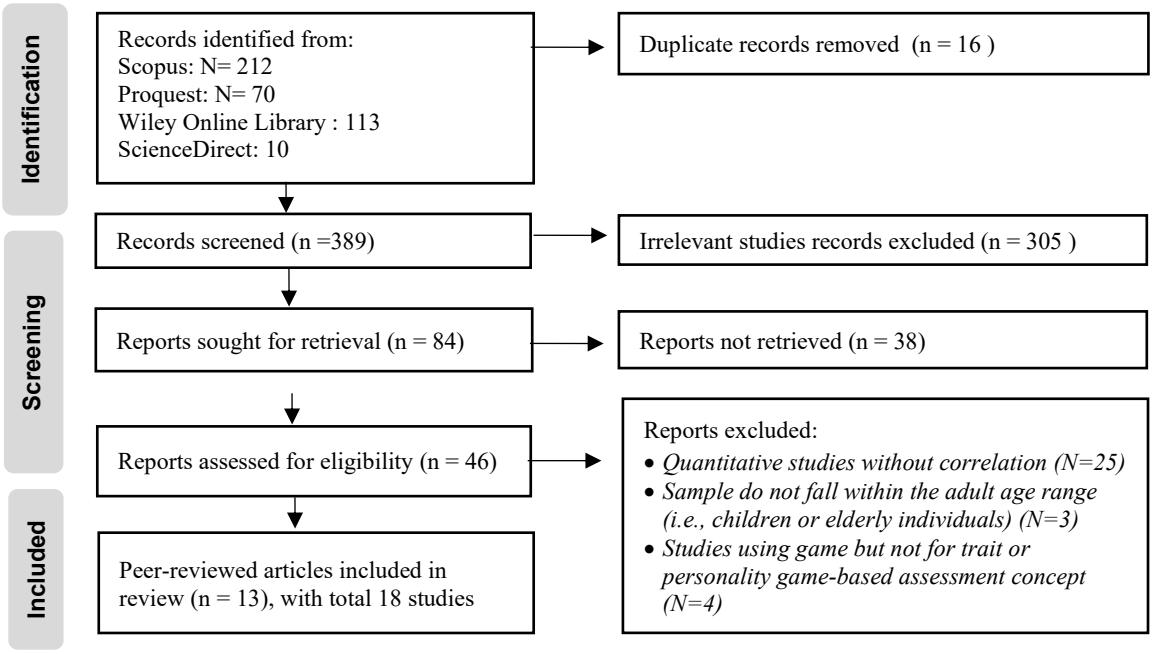

Figure 1. Review Process Based on PRISMA Guideline

2.3 Selection Process

Three researchers undertook a selection process based on inclusion and exclusion criteria independently, two researchers are doctoral candidates in psychology with research experience in assessment tools design, and one research assistant has a bachelor's degree in statistics, so it is expected that they all share a common understanding of the quantitative data required for this meta-analysis research. The initial selection involved filtering by title and abstract, with the removal of irrelevant literature. The process utilized the assistance of the reference management application Rayyan.ai. Disagreements concerning the inclusion or exclusion of an article were resolved through discussion among all reviewers, with the final decision resting with the principal investigator. Subsequently, the next step involved a full-text review of the list of potentially relevant articles.

2.4 Coding Procedure

After acquiring the relevant manuscript samples, correlation coefficients between scores from personality measures based on GBAs and from other personality scales were gathered. Additionally, the authors compiled information regarding the source of the article, the year of publication, the sample size, and the characteristics of the sample, including the average age of the participants.

2.5 Meta-Analytic Method and Statistical Analysis

The random effects model assumption was utilized because of non-identical parameters in the included studies [27]. Statistical analysis for the generalization of validity referred to an effect size index that represented the degree of correlation between the dependent and independent variables in each study [28]. The effect size involved a transformation of the correlation coefficients in each study into Fisher's Z values [29].

2.6 Heterogeneity and Moderator Analyses

The objective of synthesis was not merely to calculate a summarized effect but rather to comprehend the pattern of effects to observe the presence of heterogeneity. This study used the Q test to assess the true heterogeneity. If the expected value of Q was not equal and the p-value < 0.001, it suggested that there was significant heterogeneity among the studies, leading to the rejection of the null hypothesis that the true effect size was the same in all studies. It was suggested that the Q test should never be used as surrogates for the amount of true variance. Therefore, another analysis was used to enrich the purpose of heterogeneity analysis: the prediction interval. A narrow prediction interval suggested that the impact of the intervention was relatively consistent across populations while a wide prediction interval suggested that the impact of the intervention varied across populations. Moreover, The I2 statistic was also used to provide some context for understanding the forest plot. When I2 was low, the variance in the forest plot was mostly due to sampling error. When I2 was high, the variance in the forest plot provided a reasonable estimate for the variance of true effects.

After it had been established that GBA scores had a higher correlation in some factors, the researchers identified factors associated with the magnitude of the effect. In a meta-analysis, researchers had used regression to examine the relationship between covariates and effect size, a procedure commonly referred to as meta-regression analysis. The effect size is computed separately for each subgroup of studies, and then the each values are compared. Five study characteristics that might become potential categorical moderators as subgroups were the sample size, game type, number of attribute, study type and statistical methods being used.

2.7 Publication Bias

Publication bias analysis aimed to examine evidence of bias. Several methods, including Funnel Plot, Begg and Mazumdar's rank correlation test, Egger's Test, Orwin's method, and Duval and Tweedie's trim and fill, were used to test publication bias [27]. The first step was assessing potential bias by identifying an asymmetric graph in the funnel plot. Other statistical tests were used to quantify the captured bias. Begg and Mazumdar's rank correlation confirmed bias by assessing the correlation between standardized effect size and variances, while Egger's test used actual effect size values and precision. Orwin's method determined the number of hidden studies adjusting the overall effect. The final step was Duval and Tweedie's trim and fill procedure, providing an adjusted effect size estimate for funnel plot asymmetry. After completing these steps, bias could be categorized as insignificant, significant but trustworthy conclusions, or potentially severe enough to cast doubt on findings.

2.8 Statistical Software

All statistical analysis in this study was performed using the Comprehensive Meta-Analysis Software version 3.0 because the software provided the analysis technique needed, i.e quantitative synthesis of effect size, heterogeneity test and meta-regression analysis, and publication bias analysis.

3. Results

3.1 Convergent validity

The identities of the 18 included studies from 13 articles and the summary of convergent validity for each included studies are presented in Table 1 and Table 2. The numbering of the articles (e.g., 1–2, 4–6, 9–10, 11–12) reflects that multiple findings or analyses were drawn from a single publication. For instance, a single article may report several studies or present results across different constructs or contexts, each of which was treated as a separate data point in the analysis. To reflect this, the article is listed once in the table but corresponds to multiple entries in the Forest Plot or data synthesis. This approach is commonly used in meta-analyses when a single source contributes more than one relevant dataset. The findings of the meta-analysis are summarized in Figure 2, which displays the statistical outcomes of 18 included studies (from 13 peer-reviewed articles) along with their corresponding visual representation.

Table 1. Summary of Included Studied

No	Title	Author	Year	Journal	Self Report
1-2	Watch what I do, not what I say I do: Computer-based avatars to assess behavioral inhibition, a vulnerability factor for anxiety disorders	Myers et al.	2016	Computers in Human Behaviour [30]	AMBI questionnaire [31]
3	Gamification in employee selection: The development of a gamified assessment	Georgiou et al.	2019	International Journal of Selection and Assessment [32]	Resilience Scale [33], Adaptability Scale [34], HEXACO Personality Inventory [35]; Decision-making [36]

4-6	Game-like personality testing: An emerging mode of personality assessment	McCord et al.	2019	Personality and Individual Differences [5]	IPIP-50 [37]
7	Would you like to play? A comparison of a gamified survey with a traditional online survey method	Triantoro et al.	2019	International Journal of Information Management [38]	Big Five personality [39]
8	Reinforcing Stealth Assessment in Serious Games	Georgiadis et al.	2019	8th International Conference on Games and Learning Alliance [40]	NEO PI-R [41]
9-10	The potential of the game- and video-based assessments for social attributes: examples from practice	Leutner et al.	2021	Journal of Managerial Psychology [8]	The Big Five inventory [42]
11-12	Construct and Predictive Validity of an Assessment Game to Measure Honesty-Humility	Barends et al.	2021	Sage Journals [14]	HEXACO [43]
13	Who am I? - Development and Analysis of an Interactive 3D Game for Psychometric Testing	Afroza et al.	2021	Australasian Computer Science Week 2021 [18]	Big Five personality [44]
14	Gamifying a Personality Measure by Converting it into a Story: Convergence, Incremental Prediction, Faking, and Reactions	Landers & Collmus	2022	International Journal of Selection and Assessment [13]	IPIP NEO [37]
15	Individual differences at play: An investigation into measuring Big Five personality facets with game-based assessments	Wu et. Al.	2022	International Journal of Selection and Assessment [16]	IPIP-NEO [45]
16	Measuring Personality Through Images: Validating a Forced-Choice Image-Based Assessment of the Big Five Personality Traits	Hilliard et al.	2022	Journal of Intelligence [46]	Big Five personality [47]
17	Predictive Validity, Applicant Reactions, and Influence of Personal Characteristics of a Gamefully Designed Assessment	Ramos-Villagrasa & Fernández-Del-Río	2023	Journal of Work and Organizational Psychology [48]	Big Five personality [49]

18	Are serious games an alternative to traditional personality questionnaires? Initial analysis of a gamified assessment	Ramos-Villagrasa et al.	2024	PLoS ONE [15]	Big Five personality [49]

Table 2. Summary of Convergent Validity in Included Studies

Study No.	Study Name	N	Effect Direction	r	M (Age)	Population Characteristic
1	Myers et al. [30]	114	Positive	0,783	21,4	Undergraduate students
2	Myers et al.	210	Positive	0,780	21	Undergraduate students
3	Georgiou et al. [32]	97	Positive	0,448	26,5	Employee and job seeker
4	McCord et al. [5]	77	Positive	0,392	24	Adult from Reddit
5	McCord et al.	98	Positive	0,204	24	Undergraduate students
6	McCord et al.	338	Positive	0,272	24	Adult from Reddit
7	Triantoro et al. [38]	694	Positive	0,360	22,4	College student
8	Georgiadis [40]	80	Positive	0,782	22,4	Adult
9	Leutner et al. [8]	729	Positive	0,420	33	Adult form online panel service
10	Leutner et al.	729	Positive	0,390	33	Adult form online panel service
11	Barends et al. [14]	116	Positive	0,330	23,48	Graduates
12	Barends et al.	287	Positive	0,280	39,85	Adult form online panel service
13	Afroza et al. [18]	30	Positive	0,171	21,3	Adult
14	Landers & Collmus [13]	352	Positive	0,160	23,86	Undergraduate students
15	Wu, Mulfinger, Alexander, et al. [16]	142	Positive	0,101	20,1	psychology students
16	Hilliard et al. [46]	108	Positive	0,702	<40	Adult form online panel service
17	Ramos-Villagrasa & Fernández-Del-Río [48]	182	Positive	0,475	21,68	Undergraduate students
18	Ramos-Villagrasa et al. [15]	98	Positive	0,798	23,1	University students

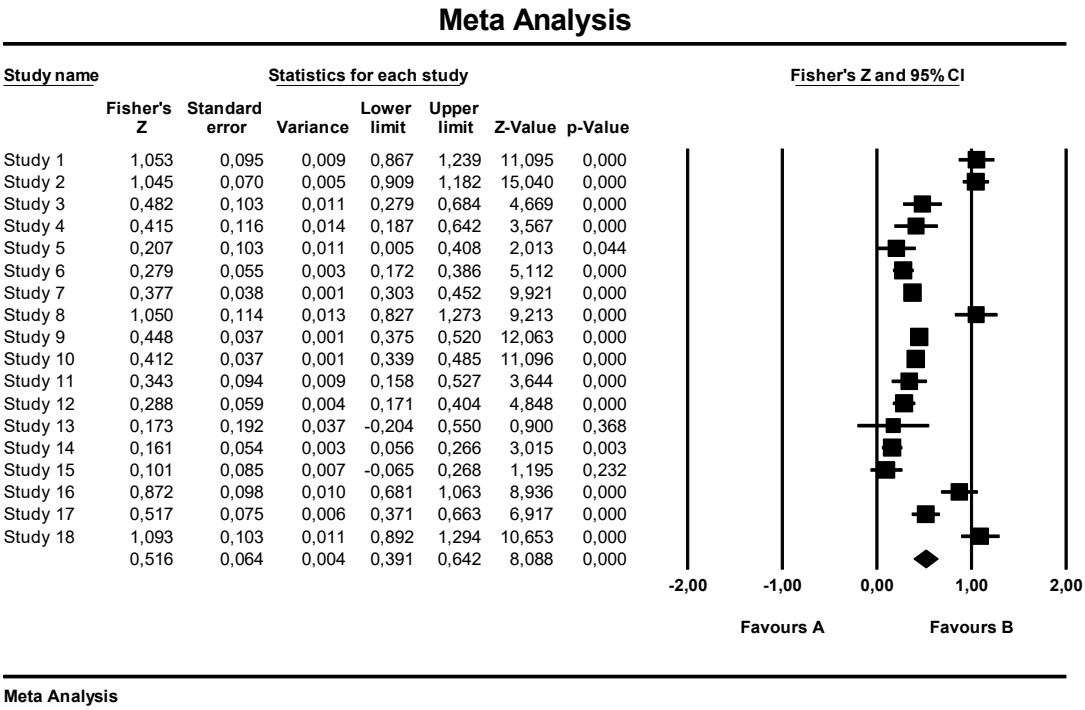

Figure 2. Forest Plot of Study Effect

Based on the Forest Plot above, the average effect size from all studies was 0,516 (95% CI: 0,391–0,642). This suggests a moderate positive relationship between GBA measurement scores and self report measures. The Z-value of 8,088 and a p-value of 0,000 indicate that the null hypothesis, asserting no association, was rejected, providing strong evidence for a non-zero average effect size in the population of comparable studies.

There was considerable variation in Fisher's values across studies, ranging from 0,101 in Study 15 to 1,093 in Study 18. Similarly, correlation coefficients ranged from 0,101 to 0,798, reflecting differences in study designs and sample characteristics. Study 9–10 reported the largest sample size (N = 729), while the Study 14 had the smallest sample size (N=30).

3.2 Heterogeneity

Table 3. Heterogeneity Test Across Studies

I^2	Q	df	p
93,868	277,240	17	0,000

The results of the meta-analysis in Table 3 reveal significant heterogeneity across the included studies. The Q statistic (Q[17] = 277,240, p = 0,000) confirms that the variation in effect sizes cannot be attributed to sampling error alone. This finding is visually represented in the forest plot, where the effect sizes span a wide range from 0,101 to 1,093, highlighting the diversity in study outcomes. The I^2 statistic further supports this interpretation, with a value of 93,868%, indicating that a substantial proportion of the observed variability is due to true differences in effect sizes rather than random error. The T^2 value of 0,0655 and T value of 0,256 provide a quantification of the estimated variance and standard deviation of the true effect sizes, offering additional insight into the extent of heterogeneity. Despite the variability, the forest plot shows that the confidence intervals for the individual studies tend to overlap, indicating a relative consistency in the intervention's overall effect.

Table 4. Recapitulation of The Moderate Variables

Study No.	Author	Group of N	Method	Game Type	Number of Atribute	Study Type
1	Myers et al. [30]	100-300	Pearson Correlation	Gamified Assessment	1	Theory driven
2	Myers et al.	100-300	Pearson Correlation	Gamified Assessment	1	Theory driven
3	Georgiou et al. [32]	<100	Regression	Gamified Assessment	4	Theory driven
4	McCord et al. [5]	<100	Pearson Correlation	Gamified Assessment	5	Theory driven
5	McCord et al.	<100	Pearson Correlation	Gamified Assessment	5	Theory driven
6	McCord et al.	>300	Pearson Correlation	Gamified Assessment	5	Theory driven
7	Triantoro et al. [38]	>300	Regression	Gamified Assessment	5	Theory driven
8	Georgiadis [40]	<100	Regression	GBA	5	Data driven
9	Leutner et al. [8]	>300	Pearson Correlation	GBA	1	Theory driven
10	Leutner et al.	>300	Pearson Correlation	GBA	1	Theory driven
11	Barends et al. [14]	100-300	Pearson Correlation	GBA	1	Theory driven
12	Barends et al.	100-300	Pearson Correlation	GBA	1	Theory driven
13	Afroza et al. [18]	<100	Pearson Correlation	GBA	3	Theory driven
14	Landers & Collmus [13]	>300	Pearson Correlation	Gamified Assessment	2	Theory driven
15	Wu, Mulfinger, Alexander, et al. [16]	100-300	Regression	GBA	5	Data driven
16	Hilliard et al. [46]	100-300	Regression	Gamified Assessment	5	Data driven
17	Ramos-Villagrasa & Fernández-Del-Río [48]	100-300	Pearson Correlation	GBA	2	Theory driven
18	Ramos-Villagrasa et al. [15]	<100	Pearson Correlation	Gamified Assessment	5	Theory driven

3.3 Moderating Variable Analysis

The analysis of moderating variables revealed several findings as refers in Table 5. For sample size, studies with 100–300 participants demonstrated the largest effect size (r = 0,602 ; p = 0,001), followed by studies with fewer than 100 participants (r = 0,579 ; p = 0.001), while studies with more than 300 participants exhibited the smallest effect size (r = 0,342 ; p = 0,000). Despite these differences, the variation in effect sizes across sample size groups was not statistically significant (Q[2] = 4,398, p = 0,111).

Table 5. Analysis Result of Moderating Variables

No	Moderator Var.	Group	n	Effect Size	Test of null (2-tail)		Heterogenity		
					z-value	P-value	Between Classes Effect (Q)	Df (Q)	p-value
1.	Sample size	<100	6	0,579	3,474	0,001	4,398	2	0,111
		100-300	7	0,602	4,118	0,000			
		>300	5	0,342	7,335	0,000			

2.	Game Type	Gamified Assessment	10	0,595	5,237	0,000	1,850	1	0,174
		GBA	8	0,419	6,370	0,000			
3.	Number of Atribute	1	6	0,593	5,292	0,000	4,514	4	0,341
		2	2	0,335	1,887	0,059			
		3	1	0,173	0,900	0,368			
		4	1	0,482	4,669	0,000			
		5	8	0,542	4,781	0,000			
4.	Study Type	Data Driven	3	0,671	2,204	0,027	0,347	1	0,557
		Theory Driven	15	0,488	7,573	0,000			
5.	Method	Pearson Correlation	13	0,498	6,477	0,000	0,182	1	0,669
		Regression	5	0,569	3,874	0,000			

Regarding game type, both Gamified Assessments (r = 0,595; p= 0,000) and Game-Based Assessments (GBA) (r = 0,419; p = 0,000) showed significant effect sizes. However, the differences in effect sizes between the two types of games were also not statistically significant (Q[1] = 1,850, p = 0,174). The analysis of the number of attributes were not statistically significant differences in effect sizes (Q[4] = 4,514; p = 0,341). Studies that measured only one attribute within a game reported the largest effect size (r = 0,593; p = 0,000), while studies that measured three attributes reported the smallest effect size (r = 0,173; p = 0,368).

For design approach, the Data-driven group exhibited a larger effect size (r = 0,671; p = 0,027) compared to the Theory-driven group (r = 0,488; p = 0,000). However, the differences in effect sizes between these two types of studies were not statistically significant (Q[1] = 0,347; p = 0,557). Lastly, for statistical methods, studies employing regression analysis showed a slightly larger effect size (r =0,569; p = 0,000) compared to those using Pearson correlation (r = 0,498; p = 0,000). Similar to other variables, the differences between these two methods were not statistically significant (Q[1] = 0,182; p = 0,669).

Overall, while certain patterns emerged in the effect sizes across moderator variables, none of the observed differences were statistically significant. This suggests that factors such as sample size, design approach, game type, number of attributes measured, and statistical methods may not play a decisive role in determining the effectiveness of game-based assessments. Although some groups exhibited larger effect sizes than others, the lack of statistical significance indicates that these variations could be due to chance rather than a meaningful impact of the moderating variables.

3.4 Publication Bias

Figure 3 displays a funnel plot where the X-axis shows the values of Fisher's Z from the reviewed studies, and the Y-axis shows their standard errors. In the plot, larger studies are positioned at the top, while smaller studies are at the bottom. The plot appears to be asymmetrical, with many studies clustered on the right side and fewer on the left. This suggests a potential publication bias, where studies with certain results might be more likely to be published.

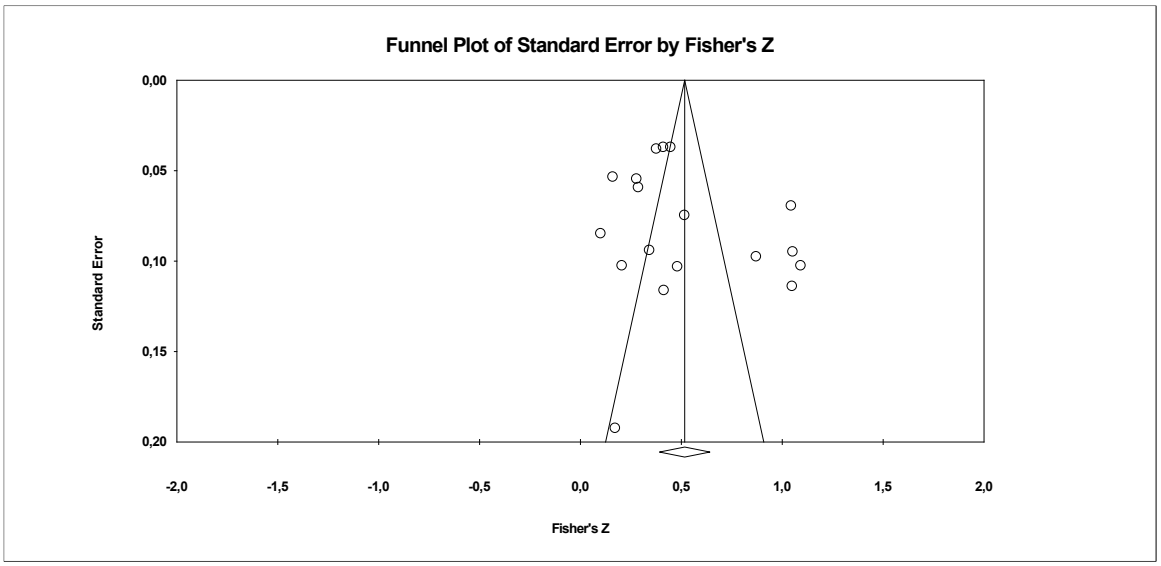

Figure 3. Funnel Plot of Standard Error

Table 6. Begg and Mazumdar Rank Correlation Test for Funnel Plot Asymmetry

Kendall's Tau	Z-value for Tau	p-value
0,163	0,947	0,344

The Begg and Mazumdar rank correlation test in Table 6 produced a Kendall's tau of 0,163, with a p-value of 0,344, indicating no strong evidence of bias.

Table 7. Egger's Regression Intercept

Intercept	Standard Error	95% CI		t-value	df	p-value
		lower limit	Upper limit			
3,07325	2,23182	-1,65800	7,80450	1,37701	16	0,18747

Similarly, Egger's test in Table 7 showed an intercept (b) of 3,07325, with a confidence interval from -1,658 to 7,805 and a p-value of 0,187, further suggesting no significant bias.

Table 8. Orwin's fail-safe N

Orwin's fail-safe N	**Criterion**
Fisher's Z in observed studies	0,445
The criterion for a trivial Fisher's Z	0,450
Mean Fisher's Z in missing studies	0,500
Number of missing studies to bring Fisher's Z over 0,490	2,000

The analysis revealed that only two missing studies with an average Fisher's Z of 0.500 would be required to increase the overall Fisher's Z beyond 0.490, which is close to the defined threshold of a trivial effect. This result suggests that the observed effect size is not highly sensitive to the inclusion of a small number of missing studies, thereby providing evidence of its robustness.

4. Discussion

4.1 Convergent validity

The results of this meta-analysis indicate an average effect size of 0.516 (95% CI: 0.391–0.642), suggesting a moderate positive relationship between GBA measurement scores and self-reported measures. While this is considered moderate based on older standards, it is regarded as fairly strong under modern benchmarks [14]. The significant Z-value of 8.088 and p-value < 0.001 provide evidence against the null hypothesis of no association, supporting the presence of a consistent effect across the studies analyzed. This finding highlights a potential link between the constructs measured, though its practical significance may vary depending on specific contexts.

Substantial variability was observed in the effect sizes across studies, with Fisher's Z-values ranging from 0.101 (Study 15) [16] to 1.093 (Study 18) [15]. Correlation coefficients similarly ranged from 0.101 to 0.798, reflecting diverse study designs, sample characteristics, and measurement approaches. Larger sample sizes, such as those in Study 9–10 (N = 729) [8], may have provided more stable estimates compared to smaller samples like that in [18] (N = 30), which could be more susceptible to variability. These differences suggest that study-specific factors may play an important role in shaping observed outcomes. In conclusion, while this meta-analysis provides evidence supporting the convergent validity of GBA metrics, further research is necessary to refine their use and clarify the contexts in which they are most effective.

4.2 Heterogeneity

The findings in this meta-analysis suggest that while GBA generally show positive effects, the degree of alignment with traditional assessments can vary significantly across studies. The significant heterogeneity in effect sizes—ranging from small to large—reflects how the effectiveness of GBA may depend on different factors such as the design of the games, the population being tested, and the specific traits or attributes being measured.

The prediction interval, which indicates potential for negative or negligible effects in some studies, implies that GBA do not universally demonstrate convergent validity with self-report assessments. This means that while GBA may show promise in certain contexts, they may not always align with traditional assessments in all situations. The overlapping confidence intervals in the forest plot suggest that, despite this variability, there is still a general trend indicating positive convergent validity, but the degree of this validity can vary.

T^2 and T values provide further insights into the variance and standard deviation of effect sizes, reinforcing the notion that the convergent validity of GBA is not uniform and may be influenced by various factors. These findings suggest that futher research on GBA should focus on identifying the specific conditions, populations, and assessment designs that lead to stronger convergent validity with traditional methods. This research will help clarify when and how GBA can be reliably used as alternatives or complements to traditional assessment tools.

4.3 Moderating Variable

The findings highlight key patterns in the influence of moderating variables on effect sizes, providing insights into the robustness and variability of outcomes under different conditions.

4.3.1 Sample Size

The analysis of moderating variables highlighted potential differences in effect sizes based on sample size categories. Studies with 100–300 participants showed the largest effect size (r = 0.602; p = 0.001), followed closely by studies with fewer than 100 participants (r = 0.579; p =

0.001). In contrast, studies with more than 300 participants exhibited the smallest effect size (r = 0.342; p = 0.000). While these differences in effect sizes suggest that smaller and medium-sized studies may report stronger associations, the statistical test for heterogeneity across sample size groups was not significant (Q[2] = 4.398, p = 0.111). This indicates that the observed variation in effect sizes may not be meaningfully attributed to sample size differences.

These findings suggest that sample size may influence effect size estimates to some extent, potentially due to methodological or contextual factors associated with smaller or medium-sized studies, such as differences in participant characteristics or research design. However, the lack of statistical significance in the moderation test implies that the relationship between sample size and effect size requires further investigation to determine whether these patterns hold consistently across diverse study contexts. Future research could explore additional moderators or examine whether there was bias or other factors contribute to these observed trends.

Previous research by [50] explained that while sample size affects the precision of estimates, the magnitude of effects in meta-analyses tends to remain stable across sample sizes if study designs are optimal. A moderate sample size may offer a balance between statistical power and reduction of unwanted variability, but these nonsignificant results may also reflect the presence of a ceiling effect, as proposed by [51], where theoretically strong relationships remain stable regardless of sample size variation.

The validity and reliability of GBA can be influenced by the sample size used in prior studies. Larger sample sizes are generally preferred as they provide more robust data and help in generalizing the findings [48]. Despite this, many studies in GBAs have been limited by small sample sizes, which can affect the strength and generalizability of the results [52]. It is suggested that future research should aim to use more diverse samples to enhance the robustness and generalizability of findings.

Research from [18] has also demonstrated that, while larger sample sizes provide more precise estimates, smaller sample sizes do not necessarily preclude meaningful findings in game-based assessments. In particular, studies with small sample sizes often focus on more homogenous participant groups, which can reduce potential variability and increase the likelihood of detecting significant effects. Furthermore, in game-based assessments, the nature of the design and the types of measurements used can influence the impact of sample size on the results.

In contrast, when studies are based on larger and more diverse samples, the results can become more generalized, allowing for better understanding of the broader applicability of the findings. As noted by [48], larger sample sizes tend to yield more robust data, particularly when it comes to generalizing findings across diverse populations, which is crucial for ensuring that game-based assessment tools are effective in a variety of settings.

These findings suggest that, while sample size remains an important consideration in the analysis of game-based design testing, the effects of sample size may be moderated by other factors, including the homogeneity of the sample and the interactive nature of the assessment. Future research should continue to explore these factors and how they interact with sample size in shaping the outcomes of game-based assessments, particularly in educational and psychological contexts.

These findings also carry important implications for the future design and validation of GBAs. Developers and researchers should consider that the strength of observed validity may be partially shaped by sample characteristics, including size, composition, and context. While larger samples are critical for ensuring generalizability and statistical precision, smaller and medium-sized samples may reveal stronger associations due to more targeted or homogenous participant groups, as well as greater control over experimental conditions. As such, future GBA research and development should adopt multi-phase validation strategies, beginning with

small, focused samples for design calibration and construct refinement, followed by large-scale implementations to assess generalizability across diverse populations.

4.3.2 Game Type

Gamified assessments have demonstrated similar reliability and validity to traditional measures, showcasing their potential as robust tools for assessing constructs like personality traits. For instance, the VASSIP gamified assessment exhibited comparable reliability and participant scoring patterns to its original personality measure [48]. Additionally, a study examining a GBA of Honesty-Humility revealed convergent validity with self-reported measures and divergent validity with unrelated traits (Barends et al., 2021). It also reported a moderate alignment between the GBA and the HEXACO model [53].

Regarding game type, both Gamified Assessments ($r = 0.595$; $p = 0.000$) and Game-Based Assessments (GBA) ($r = 0.419$; $p = 0.000$) showed significant effect sizes. However, the differences in effect sizes between the two types of games were not statistically significant ($Q[1] = 1.850$, $p = 0.174$). This suggests that both approaches are similarly effective in capturing the intended constructs, further emphasizing their robustness as assessment tools.

These results align with the theory of construct equivalence, which posits that both Gamified Assessments and GBA likely evaluate similar underlying traits, such as personality characteristics. Prior research by [9] further supports this perspective by exploring distinctions between game-based, gamified, and gamefully designed assessments in employee selection contexts. They emphasized that both Gamified Assessments and GBAs exhibit strong psychometric properties and predictive validity, measuring comparable latent constructs through distinct designs. This allows organizations to choose the most suitable format based on specific needs and feasibility. Overall, the comparable effect sizes and shared theoretical foundations of Gamified Assessments and GBAs highlight the importance of prioritizing high-quality measurement design over the choice of format.

These findings carry meaningful implications for the future development and implementation of GBAs. Given the comparable psychometric performance of gamified assessments and GBAs, developers and practitioners are encouraged to prioritize design decisions that enhance construct alignment, user engagement, and practical applicability, rather than focusing solely on format. For contexts where scalability, time efficiency, and candidate acceptance are critical, such as high-volume recruitment, gamified assessments may offer a more accessible and cost-effective solution due to their integration within existing survey-based platforms. In contrast, fully immersive GBAs may be better suited for contexts demanding richer behavioral data or where task authenticity is a priority. Importantly, developers should ensure that construct validity is not sacrificed for entertainment value, and that design choices are guided by empirical evidence rather than novelty.

4.3.3 Number of Attributes

The variation in effect sizes based on the number of attributes measured is particularly noteworthy. Studies that measured only one attribute reported the largest effect size ($r = 0.593$; $p = 0.000$), while those measuring three attributes reported the smallest effect size ($r = 0.173$; $p = 0.368$). However, the statistical analysis revealed no significant differences in effect sizes across groups ($Q[4] = 4.514$; $p = 0.341$).

Designing game elements to measure specific personality traits poses significant challenges due to the inevitable influence of other traits. For instance, in an extraversion test, an element intended to measure extraversion may inadvertently assess conscientiousness, such as when participants choose to help others, a behavior linked to a sense of responsibility. According to prior research by [18], this overlap complicates the identification of which trait primarily drives a participant's decision, ultimately impacting the accuracy of measurement outcomes.

This result aligns with cognitive load theory, which posits that as task complexity increases, the cognitive resources required to process and respond to the task also increase, potentially reducing the strength of the observed relationships [16]. Single-attribute assessments may provide more focused and precise measurements by minimizing noise and enhancing the clarity of the relationships being tested. In contrast, when multiple attributes are assessed simultaneously, overlapping constructs or competing demands can introduce additional variance, thereby weakening the overall effect size. These findings emphasize the importance of task simplicity in assessment design and suggest that overloading assessments with multiple constructs may reduce their overall effectiveness.

These findings have important implications for the future development of GBAs, particularly with respect to construct specificity and task design. The trend indicating stronger effect sizes in single-attribute assessments suggests that focused trait measurement may enhance both clarity and validity. Developers should carefully consider limiting the number of psychological attributes assessed within a single game environment to reduce cognitive load and construct interference. One practical recommendation is to design modular GBAs, where each module targets a specific trait with clearly aligned tasks and game mechanics. This modular approach would allow for sequential or adaptive deployment based on assessment needs while maintaining construct purity. Additionally, when multiple traits must be assessed within one game, integrating dynamic task-switching or branching narrative structures could help isolate specific trait expressions across contexts, reducing construct overlap.

4.3.4 Study Type

Although Data-Driven studies produced slightly larger effect sizes ($r = 0.671$; $p = 0.027$) compared to Theory-Driven studies ($r = 0.488$; $p = 0.000$), the difference was not statistically significant ($Q[1] = 0.347$; $p = 0.557$). This suggests that the methodological approach, whether data-driven or theory-driven, may not strongly influence the observed effect sizes. Theoretically, this finding could indicate that the constructs being measured are inherently robust and not overly sensitive to the approach taken.

Data-driven studies, which often employ exploratory methodologies and machine learning techniques, excel at identifying nuanced patterns within data. On the other hand, theory-driven studies emphasize a priori hypotheses and structured testing grounded in established theoretical frameworks. The similarity in effect sizes between these two approaches suggests that, as long as studies adhere to rigorous methodological standards, the outcomes remain consistent regardless of the guiding framework.

However, it is important to note that data-driven methods, while effective for assessing constructs such as cognitive ability, have shown limited evidence for modeling personality traits using trace data [54]. This indicates potential limitations of the data-driven approach for capturing certain dimensions of personality. These findings highlight the importance of selecting the most appropriate methodology for different aspects of psychological assessment to ensure accurate and reliable measurement of diverse constructs.

Given the comparable effect sizes observed between data-driven and theory-driven approaches, future GBA development should consider adopting a hybrid strategy that leverages the strengths of both methodologies. Theory-driven models can provide the conceptual clarity necessary for construct validity, especially in domains such as personality assessment where interpretability and theoretical grounding are essential. At the same time, data-driven techniques offer opportunities for enhancing predictive precision and uncovering complex behavioral patterns, particularly when applied to dynamic or context-specific traits. Researchers and practitioners should therefore consider integrating theoretical models into the initial design phase, followed by iterative refinement using empirical gameplay data. Although such a hybrid approach may be complex and not straightforward to implement, requiring multidisciplinary collaboration, large datasets, and continuous validation, it offers a promising

pathway toward more robust, adaptable, and meaningful assessment tools. This dual-phase strategy has the potential to simultaneously support construct validity and ecological validity, thereby enhancing both the scientific rigor and practical relevance of future GBAs.

4.3.5 Statistical Method

The choice of statistical method also did not significantly influence effect sizes, with regression analysis yielding slightly larger effect sizes than Pearson correlation. This consistency suggests that the methods employed to analyze data provide reliable and comparable findings. Theoretically, both regression and correlation measure relationships between variables, with regression allowing for greater flexibility in controlling for additional covariates. The lack of significant differences indicates that convergent validity index is a stable construct that is not overly dependent on the choice of analytical technique [55]. This stability aligns with theoretical expectations, as convergent validity should ideally remain robust across different statistical methods if the underlying constructs are well-defined [56].

Correlation analysis has demonstrated moderate relationships between choices in game-based personality assessments and scores on traditional Five-Factor Model inventories, supporting acceptable construct validity [5], [57]. Additionally, regression models, such as ordinary least squares (OLS) and random forests regression, have been utilized to predict scores on traditional personality and cognitive ability measures based on game assessment data, with random forests explaining a significant portion of the variance in cognitive ability prediction [16], [48]. Furthermore, machine learning techniques like Lasso regression have been recommended for scoring forced-choice, image-based personality measures in game-based assessments, due to their strong generalizability and convergent validity [46], [58]. In conclusion, both regression analysis and correlation methods produced reliable and comparable results, with regression yielding slightly larger effect sizes. This consistency suggests that convergent validity is stable across different statistical techniques, supporting the robustness of the construct. Game-based assessments demonstrated moderate correlations with traditional Five-Factor Model inventories, and regression models, including random forests and machine learning techniques like Lasso regression, effectively predicted personality and cognitive ability scores, further validating their use.

These findings carry several practical implications for the future development and application of GBAs. The observed consistency across statistical methods, particularly between correlation and regression analyses, suggests that developers and researchers can confidently apply a range of conventional analytical techniques to evaluate the validity of GBAs. For practitioners in applied settings such as human resources or educational assessment, this implies that simple correlation analyses may be sufficient for initial validation studies, especially when working with limited sample sizes or resource constraints. However, as GBAs become more sophisticated and capable of capturing complex user behaviors (e.g., reaction time, decision pathways, adaptive responses), the use of regression-based models, such as ordinary least squares (OLS) or basic machine learning methods like Lasso regression, may offer added value. These methods can accommodate multiple behavioral predictors and help improve the interpretability of how in-game actions relate to psychological traits. Nonetheless, these models should be used judiciously, with attention to overfitting, transparency in variable selection, and alignment with theoretical constructs.

For future GBA development, designers are encouraged to structure games in ways that yield analyzable data, such as clear decision points, scoring mechanics, and time-based events, so that statistical models can extract meaningful patterns. While advanced machine learning methods may enhance predictive accuracy, developers should prioritize psychometric transparency and model interpretability over algorithmic complexity, especially in high-stakes settings like employment or academic placement.

The variation in effect sizes is noteworthy in some variables, although none of the moderators, such as the number of attributes, sample size, game type, study type, and statistical methods, showed a significant influence. Despite the lack of significant differences, the observed variation highlights the potential impact of these variables on the outcomes. Future research should explore how these moderators interact and their implications for optimizing assessment and intervention designs, with particular attention to the complexity of attributes.

4.4 Publication Bias

This study also examined the potential influence of publication bias, the tendency for studies with significant results to be more likely published than those with non-significant or null findings. Several methods were employed to evaluate this possibility. The funnel plot presented in Figure 3 displayed some asymmetry, with a greater concentration of studies on the right side and fewer on the left. While such asymmetry can suggest the presence of publication bias, further statistical analyses were conducted to provide a more comprehensive assessment.

The Begg and Mazumdar rank correlation test did not indicate a meaningful association between sample size and effect size (Kendall's tau = 0.163, p = .344), suggesting that smaller studies did not systematically report stronger effects. Similarly, Egger's test produced a non-significant result (intercept = 3.07, 95% CI [-1.66, 7.80], p = .187), reinforcing the interpretation that the observed asymmetry may not be due to publication bias.

Additionally, Orwin's fail-safe N analysis indicated that only two additional studies with negligible effects would be needed to meaningfully reduce the overall effect size. This low threshold suggests that even if a small number of unpublished studies exist, their influence on the meta-analytic conclusions would likely be minimal. Taken together, while the visual pattern in the funnel plot suggests possible asymmetry, the results of the accompanying statistical tests provide no strong evidence of substantial publication bias. Thus, the overall findings of this meta-analysis appear to be robust and unlikely to be significantly distorted by selective publication.

4.5 Interpretating Convergent Validity in Context

Most of the included studies used traditional self-report instruments as the reference measures for assessing the convergent validity of GBA. This approach is understandable, given that self-reports remain the most widely available and commonly used method in personality assessment. Even though it can be reasonably assured that all included studies employed self-report instruments that were rigorously validated, this reliance nonetheless introduces methodological limitations that warrant critical reflection.

Borsboom (2005) has raised fundamental concerns about the overreliance on convergent validity as a primary validation strategy [59] . The correlation between two instruments does not necessarily imply that both assess the same underlying psychological construct. When a newly developed test is validated by correlating it with a reference measure that itself may have theoretical or psychometric shortcomings, the resulting claim of validity becomes potentially circular. Borsboom refers to this issue as a "regression to infinity," where test A is validated against test B, which was in turn validated against test C, and so on—without a clear theoretical anchor or ontological basis.

This critique is particularly relevant to the present meta-analysis, considering that the majority of reference instruments in the reviewed studies were self-report measures. Although these self-reports were rigorously validated, they remain vulnerable to inherent limitations such as social desirability bias or reliance on self-awareness. Thus, while the meta-analysis yielded statistically significant convergent relationships between GBA and self-report tools, these findings should be interpreted with appropriate caution.

Moreover, from both a technical and practical standpoint, employing another GBA as a reference standard is currently not feasible. This is due to the absence of standardized GBA frameworks, limited cross-platform compatibility, and the context-specific nature of in-game performance metrics. As a result, researchers are often constrained to rely on well-established self-report instruments as the most practical and accessible benchmark for convergent validation.

Nevertheless, the aim of this study is not to undermine the value of traditional self-report methods, but rather to provide preliminary evidence that GBAs may demonstrate comparable validity within the domain of personality assessment in organizational contexts. In this respect, convergent validity serves as a useful—though not definitive—indicator of construct overlap. Future research should explore complementary validation strategies that are both psychometrically rigorous and contextually suited to the nature of GBAs, such as the application of Item Response Theory (IRT) to dynamic in-game data or the use of predictive validity through behavioral outcomes in simulated work environments.

5. Conclusions

This meta-analysis provides empirical support for the convergent validity of GBAs, demonstrating a moderate and statistically significant correlation with traditional self-report personality measures. This finding strengthens the psychometric foundation of GBAs, positioning them as a promising method for personality assessment, particularly in applied settings such as employee selection. While the observed effect size is considered moderate by traditional benchmarks, it meets, if not exceeds, current expectations for innovative assessment methods. These findings directly address the key research question: What is the degree of convergent validity demonstrated by GBAs compared to conventional personality assessments?

However, the substantial heterogeneity observed across studies suggests that the relationship between GBA and self-report measures may be influenced by study-specific characteristics. This raises important methodological questions regarding which design features or contextual variables, such as gameplay format, user interface, or assessment environment, might moderate this relationship. Although no significant moderating effects were identified for sample size, game type, or statistical method, the variability across studies underscores the need for further exploration. Statistical tests for publication bias suggest that any such bias is unlikely to have meaningfully influenced the overall findings.

Several limitations of this analysis warrant careful consideration. The heterogeneity in effect sizes may reflect the influence of unmeasured factors related to game mechanics, participant demographics, or implementation contexts. Furthermore, while this study explored a range of potential moderators, the lack of significant findings highlights the need for more nuanced investigations to determine the specific conditions under which GBAs demonstrate optimal validity.

From a practical standpoint, these findings carry important implications for sectors such as human resources, education, and training. In organizational settings, GBAs offer an engaging and potentially less biased alternative to conventional personality questionnaires, particularly in high-volume recruitment contexts where applicant fatigue and social desirability can undermine the validity of self-report tools. For example, GBAs may be used in early-stage candidate screening to unobtrusively assess traits. In educational settings, GBAs hold potential for measuring non-cognitive domain in ways that are more immersive and context-rich than traditional tests.

However, successful implementation in these domains requires attention to several factors, including the alignment of game mechanics with the psychological constructs of interest, the standardization of scoring systems, and the integration of robust validation frameworks.

Additionally, ensuring fairness, accessibility, and user experience remains essential to prevent the introduction of new biases or usability barriers. Future research should therefore not only refine the psychometric qualities of GBAs but also expand beyond convergent validation to include predictive validity, behavioral outcomes, and cross-context generalizability, key areas for advancing beyond the current state of the art.

To guide future research and development, several practical steps can be taken. First, studies should clearly describe how the game activities are linked to the personality traits being measured. This helps ensure that what the game is testing matches the intended psychological concept. Second, researchers could try testing the same GBA in different settings or with different groups to see how well the results hold up. Third, instead of relying only on self-report questionnaires, future studies should also explore comparing GBA results with other outcomes, such as job performance. Lastly, making more detailed information about game design and scoring methods available would help others build on existing work and improve consistency across studies.

Acknowledgments

This meta-analysis process was publicly accessible on OSF at https://doi.org/10.17605/OSF.IO/B6WJM. This research did not receive any monetary or in-kind funding.

Conflicts of interest

There is no conflict of interest, whether financial or nonfinancial, were identified in this study.

References

[1] K. R. Murphy and J. L. Dzieweczynski, "Why don't measures of broad dimensions of personality perform better as predictors of job performance?," in *Human Performance*, 2005, pp. 343–357. doi: 10.1207/s15327043hup1804_2.

[2] V. J. Shute and S. Rahimi, "Review of computer-based assessment for learning in elementary and secondary education," Feb. 01, 2017, *Blackwell Publishing Ltd*. doi: 10.1111/jcal.12172.

[3] S. J. Motowidlo, A. C. Hooper, and H. L. Jackson, "Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items," *Journal of Applied Psychology*, vol. 91, no. 4, pp. 749–761, Jul. 2006, doi: 10.1037/0021-9010.91.4.749.

[4] S. J. Motowidlo and M. E. Beier, "Differentiating Specific Job Knowledge From Implicit Trait Policies in Procedural Knowledge Measured by a Situational Judgment Test," *Journal of Applied Psychology*, vol. 95, no. 2, pp. 321–333, Mar. 2010, doi: 10.1037/a0017975.

[5] J. L. McCord, J. L. Harman, and J. Purl, "Game-like personality testing: An emerging mode of personality assessment," *Pers Individ Dif*, vol. 143, pp. 95–102, Jun. 2019, doi: 10.1016/j.paid.2019.02.017.

[6] J. C. Cassady and R. E. Johnson, "Cognitive test anxiety and academic performance," *Contemp Educ Psychol*, vol. 27, no. 2, pp. 270–295, 2002, doi: 10.1006/ceps.2001.1094.

[7] 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.

[8] F. Leutner, S. C. Codreanu, J. Liff, and N. Mondragon, "The potential of game- and video-based assessments for social attributes: examples from practice," *Journal of Managerial Psychology*, vol. 36, no. 7, pp. 533–547, Oct. 2021, doi: 10.1108/JMP-01-2020-0023.

[9] R. N. Landers and D. R. Sanchez, "Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation," Mar. 01, 2022, *John Wiley and Sons Inc*. doi: 10.1111/ijsa.12376.

[10] V. J. Shute, L. Wang, S. Greiff, W. Zhao, and G. Moore, "Measuring problem solving skills via stealth assessment in an engaging video game," *Comput Human Behav*, vol. 63, pp. 106–117, Oct. 2016, doi: 10.1016/j.chb.2016.05.047.

[11] Y. J. Kim and V. J. Shute, "The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment," *Comput Educ*, vol. 87, pp. 340–356, Aug. 2015, doi: 10.1016/j.compedu.2015.07.009.

[12] R. J. Mislevy, "Evidence-centered design for simulation-based assessment.," 2013. doi: 10.7205/milmed-d-13-00213.

[13] R. N. Landers and A. B. Collmus, "Gamifying a personality measure by converting it into a story: Convergence, incremental prediction, faking, and reactions," *International Journal of Selection and Assessment*, vol. 30, no. 1, pp. 145–156, Mar. 2022, doi: 10.1111/ijsa.12373.

[14] A. J. Barends, R. E. de Vries, and M. van Vugt, "Construct and Predictive Validity of an Assessment Game to Measure Honesty–Humility," *Assessment*, vol. 29, no. 4, pp. 630–650, Jun. 2021, doi: 10.1177/1073191120985612.

[15] P. J. Ramos-Villagrasa, E. Fernández-Del-Río, R. Hermoso, and J. Cebrián, "Are serious games an alternative to traditional personality questionnaires? Initial analysis of a gamified assessment," *PLoS One*, vol. 19, no. 5 May, May 2024, doi: 10.1371/journal.pone.0302429.

[16] 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.

[17] K. Werbach and D. Hunter, "Level 5 Game Changer: Six Steps to Gamification," in *For the Win, Revised and Updated Edition*, University of Pennsylvania Press, 2020, pp. 73–88. doi: 10.9783/9781613631041-006.

[18] A. Afroza, K. Murray, B. C. Wünsche, and P. Denny, "Who am I? - Development and Analysis of an Interactive 3D Game for Psychometric Testing," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Feb. 2021. doi: 10.1145/3437378.3442688.

[19] A. M. Rosenberg *et al.*, "Quantitative mapping of human hair greying and reversal in relation to life stress," *Elife*, vol. 10, Jun. 2021, doi: 10.7554/eLife.67437.

[20] T. Chamorro-Premuzic, D. Winsborough, R. A. Sherman, and R. Hogan, "New Talent Signals: Shiny New Objects or a Brave New World?," *Ind Organ Psychol*, vol. 9, no. 3, pp. 621–640, Sep. 2016, doi: 10.1017/iop.2016.6.

[21] M. K. Mount, M. R. Barrick, and J. P. Strauss, "Validity of observer ratings of the big five personality factors.," *Journal of Applied Psychology*, vol. 79, no. 2, pp. 272–280, Apr. 1994, doi: 10.1037/0021-9010.79.2.272.

[22] M. A. Mcdaniel and N. T. Nguyen, "Situational Judgment Tests: A Review of Practice and Constructs Assessed," *International Journal of Selection and Assessment*, vol. 9, no. 1–2, pp. 103–113, Mar. 2001, doi: 10.1111/1468-2389.00167.

[23] A. P. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. Sage, Newbury Park, 2018.

[24] J. Lumsden, E. A. Edwards, N. S. Lawrence, D. Coyle, and M. R. Munafò, "Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy," *JMIR Serious Games*, vol. 4, no. 2, p. e11, Jul. 2016, doi: 10.2196/games.5888.

[25] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Wiley, 2009. doi: 10.1002/9780470743386.

[26] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *J Clin Epidemiol*, vol. 134, pp. 178–189, Jun. 2021, doi: 10.1016/j.jclinepi.2021.03.001.

[27] H. R. Rothstein, A. J. Sutton, and M. Borenstein, "Publication Bias in Meta-Analysis," in *Publication Bias in Meta-Analysis*, Wiley, 2005, pp. 1–7. doi: 10.1002/0470870168.ch1.

[28] D. C. Funder and D. J. Ozer, "Evaluating Effect Size in Psychological Research: Sense and Nonsense," *Adv Methods Pract Psychol Sci*, 2019, doi: 10.1177/2515245919847202.

[29] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, “Effect Sizes Based on Correlations,” in *Introduction to Meta-Analysis*, 2021. doi: 10.1002/9781119558378.ch6.

[30] C. E. Myers *et al.*, “Watch what I do, not what I say I do: Computer-based avatars to assess behavioral inhibition, a vulnerability factor for anxiety disorders,” *Comput Human Behav*, vol. 55, pp. 804–816, Feb. 2016, doi: 10.1016/j.chb.2015.07.067.

[31] G. Gladstone and G. Parker, “Measuring a behaviorally inhibited temperament style: Development and initial validation of new self-report measures,” *Psychiatry Res*, vol. 135, no. 2, pp. 133–143, Jun. 2005, doi: 10.1016/j.psychres.2005.03.005.

[32] K. Georgiou, A. Gouras, and I. Nikolaou, “Gamification in employee selection: The development of a gamified assessment,” *International Journal of Selection and Assessment*, vol. 27, no. 2, pp. 91–103, Jun. 2019, doi: 10.1111/ijsa.12240.

[33] G. M. Wagnild and H. M. Young, “Development and psychometric evaluation of the Resilience Scale.,” *J Nurs Meas*, vol. 1, no. 2, pp. 165–78, 1993, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7850498

[34] A. J. Martin, H. Nejad, S. Colmar, and G. A. D. Liem, “Adaptability: Conceptual and Empirical Perspectives on Responses to Change, Novelty and Uncertainty,” *Australian Journal of Guidance and Counselling*, vol. 22, no. 1, pp. 58–81, Jun. 2012, doi: 10.1017/jgc.2012.8.

[35] K. Lee and M. C. and Ashton, “Psychometric Properties of the HEXACO Personality Inventory,” *Multivariate Behav Res*, vol. 39, no. 2, pp. 329–358, Apr. 2004, doi: 10.1207/s15327906mbr3902_8.

[36] C. C. Mincemoyer and D. F. Perkins, “ Assessing decision-making skills of youth,” *Paper presented at the Forum for Family and Consumer Issues.*, vol. 8, no. 1, 2003.

[37] L. R. Goldberg *et al.*, “The international personality item pool and the future of public-domain personality measures,” *J Res Pers*, vol. 40, no. 1, pp. 84–96, Feb. 2006, doi: 10.1016/j.jrp.2005.08.007.

[38] T. Triantoro, R. Gopal, R. Benbunan-Fich, and G. Lang, “Would you like to play? A comparison of a gamified survey with a traditional online survey method,” *Int J Inf Manage*, vol. 49, pp. 242–252, Dec. 2019, doi: 10.1016/j.ijinfomgt.2019.06.001.

[39] R. R. McCrae and O. P. John, “An Introduction to the Five-Factor Model and Its Applications,” *J Pers*, vol. 60, no. 2, pp. 175–215, Jun. 1992, doi: 10.1111/j.1467-6494.1992.tb00970.x.

[40] K. Georgiadis, G. van Lankveld, K. Bahreini, and W. Westera, “Reinforcing stealth assessment in serious games,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2019, pp. 512–521. doi: 10.1007/978-3-030-34350-7_49.

[41] P. Costa and R. McCrae, “The revised NEO personality inventory (NEO-PI-R),” *The SAGE Handbook of Personality Theory and Assessment*, vol. 2, pp. 179–198, Jan. 2008, doi: 10.4135/9781849200479.n9.

[42] O. P. John and S. Srivastava, “The Big Five Trait taxonomy: History, measurement, and theoretical perspectives.,” in *Handbook of personality: Theory and research, 2nd ed.*, New York, NY, US: Guilford Press, 1999, pp. 102–138.

[43] J. M. Digman, “Personality Structure: Emergence of the Five-Factor Model,” *Annu Rev Psychol*, vol. 41, no. 1, pp. 417–440, Jan. 1990, doi: 10.1146/annurev.ps.41.020190.002221.

[44] C. Soto, “Big Five personality traits,” *The SAGE encyclopedia of lifespan human development*. Thousand Oaks, CA: Sage, pp. 240–241, Jan. 01, 2018.

[45] J. A. Johnson, “Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120,” *J Res Pers*, vol. 51, pp. 78–89, Aug. 2014, doi: 10.1016/j.jrp.2014.05.003.

[46] A. Hilliard, E. Kazim, T. Bitsakis, and F. Leutner, “Measuring Personality through Images: Validating a Forced-Choice Image-Based Assessment of the Big Five Personality Traits,” *J Intell*, vol. 10, no. 1, Mar. 2022, doi: 10.3390/jintelligence10010012.

[47] M. R. Barrick and M. K. Mount, “The Big Five Personality Dimensions and Job Performance: A Meta-Analysis,” *Pers Psychol*, vol. 44, no. 1, pp. 1–26, Mar. 1991, doi: 10.1111/j.1744-6570.1991.tb00688.x.

[48] P. J. Ramos-Villagrasa and E. Fernández-Del-Río, "Predictive Validity, Applicant Reactions, and Influence of Personal Characteristics of a Gamefully Designed Assessment," *Revista de Psicologia del Trabajo y de las Organizaciones*, vol. 39, no. 3, pp. 169–174, 2023, doi: 10.5093/JWOP2023A18.

[49] C. Soto and O. John, "The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model With 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power," *J Pers Soc Psychol*, vol. 113, pp. 117–143, Apr. 2016, doi: 10.1037/pspp0000096.

[50] J. C. Valentine, D. L. DuBois, and H. Cooper, "The Relation Between Self-Beliefs and Academic Achievement: A Meta-Analytic Review," Mar. 2004. doi: 10.1207/s15326985ep3902_3.

[51] F. L. Schmidt and J. E. Hunter, *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP : SAGE Publications, Ltd, 2015. doi: 10.4135/9781483398105.

[52] K. Aneni, I. Gomati de la Vega, M. G. Jiao, M. C. Funaro, and L. E. Fiellin, "Evaluating the validity of game-based assessments measuring cognitive function among children and adolescents: A systematic review and meta-analysis," in *Progress in Brain Research*, vol. 279, F. H. Santos, Ed., Elsevier, 2023, pp. 1–36. doi: 10.1016/bs.pbr.2023.02.002.

[53] I. Nikolaou, K. Georgiou, and V. Kotsasarlidou, "Exploring the Relationship of a Gamified Assessment with Performance," *Spanish Journal of Psychology*, vol. 22, 2019, doi: 10.1017/sjp.2019.5.

[54] E. Auer, G. Mersy, S. Marin, J. Blaik, and R. Landers, "Using machine learning to model trace behavioral data from a game-based assessment," *International Journal of Selection and Assessment*, vol. 30, Dec. 2022, doi: 10.1111/ijsa.12363.

[55] T. Raykov, "Evaluation of convergent and discriminant validity with multitrait-multimethod correlations," *British Journal of Mathematical and Statistical Psychology*, vol. 64, no. 1, pp. 38–52, Feb. 2011, doi: 10.1348/000711009X478616.

[56] R. Willink, "On the validity of methods of uncertainty evaluation," *Metrologia*, vol. 47, no. 1, pp. 80–89, 2010, doi: 10.1088/0026-1394/47/1/009.

[57] J. Harman and J. Purl, "Advances in Game-Like Personality Assessment," *Trends in Psychology*, vol. 32, pp. 1–15, Mar. 2022, doi: 10.1007/s43076-022-00162-x.

[58] R. Levy, "Dynamic Bayesian Network Modeling of Game-Based Diagnostic Assessments," *Multivariate Behav Res*, vol. 54, no. 6, pp. 771–794, Nov. 2019, doi: 10.1080/00273171.2019.1590794.

[59] D. Borsboom, *Measuring the Mind*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511490026.