Validity and Reliability of Digital Game Assessments

Digital game-based assessments show moderate to strong validity (r = 0.30-0.69) and high reliability (> 0.70) compared to traditional methods, while offering improved emotional outcomes and practical advantages.

Abstract

Digital game-based assessments of cognitive variables exhibit moderate convergent validity compared with traditional tasks. In eight studies that paired the two methods, correlations ranged from 0.30 to 0.69. One study using an adaptive digital task detected attention deficits with greater sensitivity than conventional tests, and several reports noted reliability metrics—such as Cronbach's values above 0.70 and between-person reliability as high as 0.97—that support sound measurement properties. Across tests of memory, cognitive control, and executive functions, digital and traditional approaches produced comparable statistical associations.

For emotional variables, two studies provided evidence. One observed that a game-based math test reduced test anxiety and enhanced engagement relative to a paper test, while another found that virtual reality tasks were perceived as more pleasant than paper-and-pencil assessments. Additional practical advantages included faster administration (up to five times quicker) and enhanced ecological validity in certain settings.

Paper search

Using your research question ”What is the validity and reliability of digital game-based assessments compared to traditional psychological measurement tools in measuring cognitive and emotional variables?”, we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 50 papers most relevant to the query.

Screening

We screened in papers that met these criteria:

- **Comparative Assessment**: Does the study compare digital game-based assessments with validated traditional psychological measurement tools?
- **Outcome Variables**: Does the study measure cognitive variables (e.g., attention, memory, executive function) and/or emotional variables (e.g., anxiety, depression, emotional regulation)?
- **Psychometric Data**: Does the study report psychometric properties (validity and/or reliability metrics)?
- **Study Design**: Is the study either a primary research study (experimental, quasi-experimental, validation study) or a systematic review/meta-analysis containing empirical data?
- **Comparison Measure Quality**: Are the traditional assessment tools used as comparison measures standardized and validated?
- **Assessment Focus**: Is the primary focus of the study on assessment rather than intervention?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Study Design Type**:

  Identify the primary study design type from the full text. Options include:

- Validation study
- Pilot study
- Cross-sectional study
- Comparative study

Look in the methods section for explicit design description. If multiple design elements are present, select the most prominent type. If unclear, note ”design not clearly specified” and provide a brief explanation.

- **Digital Game-Based Assessment Characteristics**:

  Extract detailed information about the digital game-based assessment:

- Name of the game/assessment tool
- Specific cognitive or emotional variables being measured
- Game mechanics (e.g., adaptive, mini-games, stealth assessment)
- Platform or technology used

Locate this information in the methods or introduction sections. Be precise about game features and measurement approach. If multiple aspects are measured, list all.

- **Participant Demographics**:

  Record the following participant details:

- Total sample size
- Age range
- Gender distribution
- Any specific population characteristics (e.g., neurodevelopmental conditions)

Extract from methods section. If percentages are provided, include those. If ranges are given, specify both minimum and maximum values. If subgroups exist, break down demographics for each.

- **Comparison Measurement Tools**:

  Identify traditional psychological measurement tools used for comparison:

- Name of comparison tool
- Specific cognitive/emotional variable measured
- Standardization status of comparison tool

Look in methods and results sections. If multiple comparison tools are used, list all. Note any statistically significant differences between game-based and traditional assessments.

- **Validity and Reliability Metrics**:

  Extract quantitative validity and reliability indicators:

- Correlation coefficients
- Reliability scores (e.g., Cronbach's alpha)
- Construct validity measures
- Predictive validity indicators

Locate in results section. Include exact numerical values and statistical significance levels. If multiple metrics are reported, include all relevant data.

- **Primary Outcomes and Conclusions**:

  Summarize the key findings regarding digital game-based assessment:

- Main conclusions about assessment validity
- Significant differences from traditional methods
- Potential advantages or limitations

Extract from results and discussion sections. Focus on direct statements about the assessment's effectiveness. Include direct quotes if they succinctly capture the primary findings.

Results

Characteristics of Included Studies

Study	Study Design	Assessment Type	Variables Measured	Sample Characteristics	Full text retrieved
Aneni et al., 2023	Systematic review and meta-analysis	Game-based and traditional	Cognitive functions across neurocognitive domains	No mention found	No
Anguera et al., 2016	Pilot study	Digital game-based (EVO) and traditional (Flanker, Visual Search)	Cognitive control abilities, selective attention	111 children (20 with 16p11.2 deletion, 16 siblings, 75 neurotypical)	Yes
Bipp et al., 2024	Meta-analysis	Game-related and traditional	Cognitive ability	Over 6,100 adult participants	No
Kiili and Ketamo, 2018	Validation study	Game-based (Semideus Exam) and paper-based	Conceptual fraction knowledge, test anxiety, flow experience	51 Finnish sixth graders	No

Study	Study Design	Assessment Type	Variables Measured	Sample Characteristics	Full text retrieved
Kourtesis et al., 2020	Validation study	Virtual reality (VR-EAL) and paper-and-pencil	Prospective memory, episodic memory, attention, executive functions	41 participants (21 females, 18 gamers, 23 non-gamers)	No
Pedersen et al., 2020	Validation study	Game-based (Skill Lab) and traditional tasks	Broad suite of cognitive abilities	10,725 participants (49% female, 50% male, 1% other), aged 16 and above	Yes
Shute et al., 2016	Validation study	Game-based (Use Your Brainz) and traditional (Raven's Progressive Matrices, MicroDYN)	Problem-solving skills	47 7th grade students (20 male, 27 female)	Yes
Sliwinski et al., 2018	Validation study	Ambulatory cognitive assessments and traditional in-lab tasks	Working memory, perceptual speed	219 adults (34% men, 66% women), aged 25-65	Yes
Song et al., 2020	Validation study	Mobile game-based (CoCon) and traditional neuropsychological tests	Cognitive control (sustained attention, working memory, inhibition, categorization)	100 children and adolescents (59% male, 41% female), aged 9-16	Yes
Weiner and Sanchez, 2020	Validation study	VR game-based and traditional self-report	Space Visualization, Visual Speed & Accuracy, Visual Pursuit	124 students (71% female), mean age 24 years	Yes

Of the 10 studies we examined:

- 7 used validation study designs

- 1 was a systematic review and meta-analysis
- 1 was a meta-analysis
- 1 was a pilot study

We found that:

- 8 studies used game-based assessments
- 8 studies used traditional assessments
- 2 studies used virtual reality assessments
- 2 studies used paper-based assessments
- 1 study used ambulatory assessment

We found a range of cognitive variables measured across the studies:

- Memory measured in 4 studies
- Cognitive functions/abilities measured in 3 studies
- Attention measured in 3 studies
- Cognitive control measured in 2 studies
- Other variables included problem-solving, visual skills, executive functions, fraction knowledge, anxiety, flow experience, inhibition, categorization, and perceptual speed

We found that 8 out of 10 studies included both game-based and traditional assessment methods, allowing for comparison between these approaches.

Validity and Reliability Evidence

Cognitive Assessment Comparisons

Study	Assessment Method	Validity Coefficients	Reliability Metrics	Key Findings
Aneni et al., 2023	Game-based vs traditional	Correlation coefficient (r) = 0.3-0.69 for 75% of correlations	No mention found	Game-based assessments showed significant correlations with traditional assessments, mostly low to medium strength
Anguera et al., 2016	EVO vs Flanker and Visual Search tasks	No mention found	No mention found	EVO more sensitive in detecting cognitive deficits than traditional methods

Study	Assessment Method	Validity Coefficients	Reliability Metrics	Key Findings
Bipp et al., 2024	Game-related vs traditional	Correlation coefficient (r) = 0.30 (corrected r = 0.45)	No mention found	Game-related assessments correlated with traditional cognitive measures
Kiili and Ketamo, 2018	Semideus Exam vs paper-based test	Significant correlation reported (value not provided)	No mention found	Game-based math test scores correlated significantly with paper-based test scores
Kourtesis et al., 2020	Virtual Reality Everyday Assessment Lab (VR-EAL) vs paper-and-pencil tests	Significant correlation reported (value not provided)	No mention found	VR-EAL scores significantly correlated with paper-and-pencil test scores
Pedersen et al., 2020	Skill Lab vs traditional tasks	Out-of-sample prediction strength (rcv) > 0.2 for accepted models	Cronbach's > 0.7 for most measures	Game-based measures showed good convergent validity with traditional tasks
Shute et al., 2016	Use Your Brainz vs Raven's and MicroDYN	Correlation coefficient (r) = 0.40-0.41, $p < 0.01$	Cronbach's = 0.76 for problem-solving assessment	Stealth assessment correlated significantly with external measures
Sliwinski et al., 2018	Ambulatory vs in-lab tasks	Correlation coefficient (r) = 0.24 to 0.74	Between-person reliability 0.97, Within-person reliability 0.41-0.53	Ambulatory assessments showed high between-person reliability and moderate within-person reliability
Song et al., 2020	CoCon vs traditional tests	Correlation coefficient (r) = 0.304 to 0.483, $p < 0.05$	Cronbach's = 0.72 to 0.897 for traditional tests	CoCon was reliable and valid for assessing cognitive control
Weiner and Sanchez, 2020	VR games vs self-report assessments	Varied by cognitive ability (not all significant)	Cronbach's = 0.86 to 0.91 for self-report assessments	Mixed validity results across different cognitive abilities

We found that:

- 8 out of 10 studies compared game-based assessments to traditional methods
- 1 study compared ambulatory to in-lab tasks
- 1 study compared game-based to self-report assessments

Regarding validity:

- 8 out of 10 studies reported on validity
- 6 provided specific correlation coefficient values
- 2 reported significance without specific values
- We didn't find validity information for 1 study
- 1 study reported mixed results across different cognitive abilities

Regarding reliability:

- 5 out of 10 studies reported reliability metrics
- We didn't find reliability information for the other 5 studies

Key findings varied across studies:

- 5 out of 10 studies reported significant correlations between game-based and traditional assessments
- 1 study found game-based assessments to be more sensitive in detecting cognitive deficits
- 1 study reported good convergent validity with traditional tasks
- 1 study showed high reliability for the game-based assessment
- 1 study concluded that their game-based assessment was both reliable and valid
- 1 study reported mixed results across different cognitive abilities

Emotional/Affective Assessment Comparisons

Study	Assessment Method	Validity Coefficients	Reliability Metrics	Key Findings
Aneni et al., 2023	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Anguera et al., 2016	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Bipp et al., 2024	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Kiili and Ketamo, 2018	Game-based vs traditional for test anxiety and flow experience	No mention found	No mention found	Game-based assessment lowered test anxiety and increased engagement

Study	Assessment Method	Validity Coefficients	Reliability Metrics	Key Findings
Kourtesis et al., 2020	Virtual Reality Everyday Assessment Lab (VR-EAL) vs paper-and-pencil for user experience	No mention found	No mention found	VR-EAL tasks were more pleasant than paper-and-pencil tests
Pedersen et al., 2020	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Shute et al., 2016	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Sliwinski et al., 2018	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Song et al., 2020	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables
Weiner and Sanchez, 2020	Not applicable	Not applicable	Not applicable	We didn't find mention of emotional/affective variables

We found that:

- 2 out of 10 studies assessed emotional or affective variables
- Of these 2 studies:
  - One compared game-based assessment to traditional assessment for test anxiety and flow experience
  - One compared VR-EAL to paper-and-pencil tests for user experience
- We didn't find validity or reliability metrics reported for either of these studies
- Key findings from these 2 studies included:
  - Game-based assessment lowered test anxiety and increased engagement in one study
  - VR-EAL tasks were reported as more pleasant than paper-and-pencil tests in the other study
- We didn't find assessments of emotional or affective variables in the remaining 8 studies

Comparative Effects

Measurement Accuracy and Precision

The included studies provide mixed evidence regarding the measurement accuracy and precision of digital game-based assessments compared to traditional methods:

- Improved accuracy:
    - Anguera et al. (2016) found their digital game-based assessment (EVO) more sensitive in detecting cognitive deficits related to attention in children with 16p11.2 deletion compared to traditional non-adaptive assessments
    - Pedersen et al. (2020) reported game-based measures (Skill Lab) showed good convergent validity with traditional tasks, with accepted models having out-of-sample prediction strength (rcv) greater than 0.2
    - Shute et al. (2016) found significant correlations ($r = 0.40\text{-}0.41$) between their game-based stealth assessment and established measures of problem-solving skills
- Variable accuracy:
    - Weiner and Sanchez (2020) reported mixed validity results for their VR game-based assessments across different cognitive abilities
    - Aneni et al. (2023) found game-based assessments showed significant correlations with traditional assessments, but many correlations were in the low to medium range ($r = 0.3\text{-}0.69$)

User Experience and Assessment Bias

Several studies highlighted potential advantages of digital game-based assessments in terms of user experience and reduced assessment bias:

- Improved user experience:
    - Kiili and Ketamo (2018) found their game-based math test lowered test anxiety and increased engagement compared to traditional paper-based tests
    - Kourtesis et al. (2020) reported participants found their VR-based assessment tasks more pleasant than traditional paper-and-pencil tests
    - Pedersen et al. (2020) noted their game-based assessment (Skill Lab) provided a more engaging and scalable approach compared to traditional methods
- Potential sources of bias:
    - Aneni et al. (2023) found factors such as age, gender, and prior gaming experience may influence the validity of game-based assessments
    - Weiner and Sanchez (2020) reported their VR game-based assessments showed adverse impact across demographic groups, similar to traditional assessments

Implementation Considerations

The studies highlight several important considerations for implementing digital game-based assessments:

1. Efficiency : Pedersen et al. (2020) reported game-based measures were five times faster than equivalent task-based measures.

2. Ecological Validity : Kourtesis et al. (2020) found VR-based assessment offered enhanced ecological validity compared to paper-and-pencil tests.
3. Adaptability : Anguera et al. (2016) used adaptive algorithms in their game-based assessment.
4. Scalability : Pedersen et al. (2020) demonstrated large-scale implementation with over 10,000 participants.
5. Technological Requirements : Studies using VR or mobile platforms highlight need for specific infrastructure.
6. Design Considerations : Aneni et al. (2023) found more valid assessments measured multiple neurocognitive domains and used prediction models for scoring.
7. Population Specificity : Several studies focused on specific age groups or populations.

References

E. J. Weiner, and Diana R. Sanchez. “Cognitive Ability in Virtual Reality: Validity Evidence for Vr Game-Based Assessments.” *International Journal of Selection and Assessment*, 2020.

Hyunjoo Song, Do-Joon Yi, and Hae-Jeong Park. “Validation of a Mobile Game-Based Assessment of Cognitive Control Among Children and Adolescents.” *PLoS ONE*, 2020.

JA Anguera, AN Brandes-Aitken, CE Rolle, SN Skinner, SS Desai, JD Bower, WE Martucci, WK Chung, EH Sherr, and EJ Marco. “Characterizing Cognitive Control Abilities in Children with 16p11.2 Deletion Using Adaptive ‘Video Game' Technology: A Pilot Study.” *Translational Psychiatry*, 2016.

K. Kiili, and H. Ketamo. “Evaluating Cognitive and Affective Outcomes of a Digital Game-Based Math Test.” *IEEE Transactions on Learning Technologies*, 2018.

Kammarauche Aneni, Isabella Gomati de la Vega, Megan G. Jiao, Melissa C Funaro, and Lynn E. Fiellin. “Evaluating the Validity of Game-Based Assessments Measuring Cognitive Function Among Children and Adolescents: A Systematic Review and Meta-Analysis.” *Progress in Brain Research*, 2023.

M. K. Pedersen, Carlos Mauricio Castano D'iaz, Qian Janice Wang, Mario Alejandro Alba-Marrugo, A. Amidi, R. Basaiawmoit, Carsten Bergenholtz, et al. “Measuring Cognitive Abilities in the Wild: Validating a Population-Scale Game-Based Cognitive Assessment.” *Cognitive Sciences*, 2020.

M. Sliwinski, J. Mogle, Jinshil Hyun, Elizabeth Munoz, J. Smyth, and R. Lipton. “Reliability and Validity of Ambulatory Cognitive Assessments.” *Assessment (Odessa, Fla.)*, 2018.

Panagiotis Kourtesis, S. Collina, Leonidas A. A. Doumas, and Sarah E. MacPherson. “Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An Immersive Virtual Reality Neuropsychological Battery with Enhanced Ecological Validity.” *Journal of the International Neuropsychological Society*, 2020.

T. Bipp, Serena Wee, Marvin Walczok, and Laura Hansal. “The Relationship Between Game-Related Assessment and Traditional Measures of Cognitive Ability—A Meta-Analysis.” *Journal of Intelligence*, 2024.

V. Shute, Lubin Wang, Samuel Greiff, Weinan Zhao, and G. Moore. “Measuring Problem Solving Skills via Stealth Assessment in an Engaging Video Game.” *Computers in Human Behavior*, 2016.