Validity and Reliability of Digital Game Assessments

Digital game-based assessments show moderate to strong validity ($r = 0.30 - 0.69$) and high reliability ($\alpha > 0.70$) 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 α 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.

Methods

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

Study	Study Design	Assessment Type	Variables Measured	Sample Characteristics	Full text retrieved
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
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

Study	Study Design	Assessment Type	Variables Measured	Sample Characteristics	Full text retrieved
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.

Cognitive Assessment Comparisons

Col1	Col2	Col3	Col4	Col5
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
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 $\alpha > 0.7$ for most measures	Game-based measures showed good convergent validity with traditional tasks

Col1	Col2	Col3	Col4	Col5
Shute et al., 2016	Use Your Brainz vs Raven's and MicroDYN	Correlation coefficient (r) = 0.40-0.41, $p < 0.01$	Cronbach's $\alpha = 0.76$ 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 $\alpha = 0.72$ 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 $\alpha = 0.86$ 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

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. [2023] 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 = 0.40-0.41) 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 = 0.3-0.69)

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

Literatura

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*, pages 1–36. Elsevier, 2023. ISBN 9780443223709. doi: 10.1016/bs.pbr.2023.02.002. URL http://dx.doi.org/10.1016/bs.pbr.2023.02.002.

J A Anguera, A N Brandes-Aitken, C E Rolle, S N Skinner, S S Desai, J D Bower, W E Martucci, W K Chung, E H Sherr, and E J Marco. Characterizing cognitive control abilities in children with 16p11.2 deletion using adaptive 'video game' technology: a pilot study. *Translational Psychiatry*, 6(9):e893–e893, sep 20 2016. ISSN 2158-3188. doi: 10.1038/tp.2016.178. URL http://dx.doi.org/10.1038/tp.2016.178.

Kristian Kiili and Harri Ketamo. Evaluating Cognitive and Affective Outcomes of a Digital Game-Based Math Test. *IEEE Transactions on Learning Technologies*, 11(2):255–263, apr 1 2018. ISSN 1939-1382. doi: 10.1109/tlt.2017.2687458. URL http://dx.doi.org/10.1109/TLT.2017.2687458.

Panagiotis Kourtesis, Simona 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*, 27(2):181–196, aug 10 2020. ISSN 1355-6177. doi: 10.1017/s1355617720000764. URL http://dx.doi.org/10.1017/S1355617720000764.

Mads Kock Pedersen, Carlos Mauricio Castaño Díaz, Qian Janice Wang, Mario Alejandro AlbaMarrugo, Ali Amidi, Rajiv V. Basaiawmoit, Carsten Bergenholtz, Morten H. Christiansen, Miroslav Gajdacz, Ralph Hertwig, Byurakn Ishkhanyan, Kim Klyver, Nicolai Ladegaard, Kim Mathiasen, Christine Parsons, Janet Rafner, Anders R. Villadsen, Mikkel Wallentin, Blanka Zana, and Jacob F. Sherson. Measuring Cognitive Abilities in the Wild: Validating a PopulationScale GameBased Cognitive Assessment. *Cognitive Science*, 47(6), 6 2023. ISSN 0364-0213. doi: 10.1111/cogs.13308. URL http://dx.doi.org/10.1111/cogs.13308.

Valerie J. Shute, Lubin Wang, Samuel Greiff, Weinan Zhao, and Gregory Moore. Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63:106–117, 10 2016. ISSN 0747-5632. doi: 10.1016/j.chb.2016.05.047. URL http://dx.doi.org/10.1016/j.chb.2016.05.047.

Erik J. Weiner and Diana R. Sanchez. Cognitive ability in virtual reality: Validity evidence for VR gamebased assessments. *International Journal of Selection and Assessment*, 28(3):215–235, jun 14 2020. ISSN 0965-075X. doi: 10.1111/ijsa.12295. URL http://dx.doi.org/10.1111/ijsa.12295.