Game-Based Psychological Assessment: Concept, Paradigm, and Practice

Authors: Xu Junyi, Li Zhongquan, Li Zhongquan

Date: 2020-10-26T00:00:00+00:00

Abstract

Game-based psychological assessment refers to the quantitative evaluation of an individual' s abilities, personality, and other psychological characteristics and behaviors through games or gamified activities. Initially developed primarily for evaluating educational and training effectiveness, it has since evolved into an assessment of psychological traits. As an emerging technology, game-based assessment offers advantages in assessment format, process, and outcomes. Currently, it has established a paradigm based on Evidence-Centered Design to guide the development of assessment tools and empirical research, with practical implementations in evaluating both cognitive and non-cognitive abilities. However, the technology remains in its nascent stage, and future research could be further expanded and deepened in task design, results analysis, and practical applications.

Full Text

Preamble

Game-Based Psychological Assessment: Concepts, Paradigms, and Practices

XU Junyi, LI Zhongquan
(Department of Psychology, School of Social and Behavioral Sciences, Nanjing University, Nanjing 210023, China)

Abstract

Game-based psychological assessment refers to the quantitative evaluation of an individual' s abilities, personality, and other psychological characteristics through games or gamified activities. Initially developed primarily for assessing educational and training outcomes, it has since evolved into a technology for

evaluating psychological traits. As a novel approach, game-based assessment offers distinct advantages in terms of assessment format, process, and results. Currently, a paradigm grounded in evidence-centered design has emerged to guide the development of assessment tools and empirical research, with practical applications in evaluating both cognitive and non-cognitive abilities. However, this technology remains in its infancy, and future research could further expand and deepen investigations into task design, data analysis, and practical applications.

Keywords: game-based assessment, evidence-centered design, cognitive ability, non-cognitive ability

According to the "Statistical Report on Internet Development in China (2019)" published by the China Internet Network Information Center, as of June 2019, the number of online gamers in China reached 494 million, with users concentrated primarily between ages 10 and 39. Gaming has become an integral component of human social behavior. While previous research focused on the impact of games on individual psychology and behavior, the advent of the big data era has sparked a quiet revolution across various domains, offering new avenues for psychological research. Unlike traditional behavioral measurement methods, human behavior can now be quantified to some extent through data in the big data era (Schoedel et al., 2018). Consequently, recent years have witnessed growing interest in leveraging games to obtain rich data and predict players' knowledge, skills, and traits.

The commercial sector has begun experimenting with integrating game elements into psychological testing for corporate recruitment (Yang & Sun, 2015), developing a series of game-based systems. Arctic Shore, a pioneer in gamified recruitment, was among the first to use behavioral tasks to evaluate candidates' personality traits. PwC collaborated with them to establish the Career Unlocked gamified recruitment system, which has been implemented and includes multiple game tasks such as balloon inflation for savings and emotion recognition. Deloitte also developed its own assessment mobile game, Firefly Freedom, which evaluates candidates' traits like diligence, perfectionism tendencies, and risk management and avoidance across multiple levels.

Game-based psychological assessment expands the toolkit of psychological measurement, and its integration with machine learning methods endows it with tremendous potential in the big data era. However, as an emerging interdisciplinary field, game-based assessment currently finds itself in a stage of "practice preceding theory." The technology has yet to receive widespread attention from domestic scholars. Therefore, this paper primarily builds upon existing international research, supplemented by a small amount of domestic work, to introduce this new technology of game-based assessment. We review its concepts, assessment paradigms, and practical applications, and propose directions for future research, hoping to provide a reference for subsequent studies.

2.1 Concept Development and Definition

Games, by their nature, represent a form of entertainment involving interactive participation, with entertainment being their essential characteristic (Wu, 2015). Since the 21st century, scholars have recognized the enormous potential of games, extending their purpose beyond mere entertainment and giving rise to the concept of serious games both domestically and internationally. Serious games utilize the entertaining form of games to achieve serious purposes such as education, training, and therapy (Gamberini et al., 2009). Game-based learning, as a branch of serious games, aims to help people acquire knowledge and behaviors through games (Gee, 2008). Early game-based assessments primarily focused on evaluating specific learning outcomes and skills, developing assessment models within game environments (Mislevy et al., 2012). As the field advanced, researchers began linking games to ability traits, allowing games to provide clues about how individuals think and act.

Heinzen et al. (2015) defined game-based assessment (GBA) as evaluating a target object through games or gamified activities. From a psychological perspective, this refers to using game-based approaches to quantitatively assess an individual' s abilities, personality, and other psychological characteristics and behaviors (Sun, Li, & Fu, 2018). Based on assessment format, it can be divided into two categories: external assessment and internal assessment. On one hand, assessment can be based on evidence external to the game, such as an individual' s final solution, reasons articulated in oral statements or self-reports (Caballero-Hernández et al., 2017). On the other hand, assessment can be embedded within the game as part of the experience, also known as stealth assessment—valid tests embedded within games to measure individual performance in the game environment (Shute, 2011).

The concepts of game-based assessment and gamification in assessment are quite similar, both representing ways to apply game mechanics to non-game contexts (Attali & Arieli-Attali, 2015). However, the most important distinction lies in the purpose of introducing game mechanics. Gamification in assessment operates on the premise that games generate positive internal motivation, fostering long-term interaction between individuals and specific environments, enhancing engagement and acceptance, and positively influencing assessment performance. Therefore, gamification in assessment is primarily used in educational settings to create a favorable environment. In contrast, game-based assessment aims to measure and evaluate one or more traits of an individual based on behavioral performance data within the game, with the focus squarely on achieving assessment objectives.

2.2 Evaluation of Advantages and Disadvantages

In recent years, game-based assessment has become a popular evaluation method because, compared to traditional psychological assessment, it offers certain advantages in assessment format, process, and outcomes.

First, regarding assessment format, it creates a relatively high-fidelity environment that can measure individuals' application of knowledge and skills through complex tasks (Shute et al., 2016). Unlike traditional ability tests that rely on recognition, recall of information, or self-report methods, game-based assessment can present scenarios that give students opportunities to demonstrate their understanding and application of knowledge. Additionally, it can incorporate multiple levels to examine individual performance across different categories or difficulty levels, offering greater flexibility.

Second, regarding the assessment process, game-based assessment can reduce anxiety during testing and increase engagement, yielding more authentic data from test-takers. Compared to traditional psychological assessments that suffer from test anxiety and social desirability issues, numerous studies have shown that test-takers find game-based assessment more engaging and enjoyable (DeRosier & Thomas, 2019; Turan & Meral, 2018). Moreover, researchers have embedded assessment items within games for student testing and found that game-based assessment reduced test anxiety and significantly improved test performance (Mavridis & Tsiatsos, 2017). Furthermore, the covert nature of game-based assessment prevents test-takers from guessing the test' s intent, effectively reducing response bias.

Third, regarding assessment outcomes, game-based assessment represents a dynamic and continuous process that can track test-takers'performance throughout gameplay through computer-based process data tracking technology, whereas traditional psychological tests only yield final scores. By integrating with machine learning methods such as Bayesian networks, dynamic models can be established and updated based on test-taker performance to obtain more accurate data (Shute et al., 2016).

However, game-based assessment inevitably has some drawbacks. For researchers and users alike, significant challenges exist in three areas: assessment game development, data analysis, and result validity.

Regarding assessment game development, it typically requires integrating game mechanics, content, and assessment components, involving collaboration among researchers, game designers, and educators to produce a specialized game. This approach demands substantial investment in time, money, and human resources. Early studies often utilized existing commercial games such as Plants vs. Zombies, but these games were not originally developed to assess specific psychological traits, limiting assessment to relatively narrow topics and resulting in inaccurate and incomplete content. Some researchers have attempted to design generic game frameworks that separate content from game mechanics to develop topic-specific games and lower development barriers, but issues such as interrupted flow experiences and practice effects remain (Baron, 2017).

Regarding assessment data analysis, game-based assessment generates vast amounts of process data, such as mouse clicks and reaction times. On one hand, recording, processing, and analyzing these data are far more complex than

data from traditional psychological tests, placing high demands on researchers' analytical capabilities. On the other hand, the key to game-based assessment lies in establishing relationships between process data and the structure of the traits being measured. Determining and validating causal relationships between data indicators and target traits among numerous data points presents considerable difficulty for researchers (Kim & Ifenthaler, 2019).

Regarding assessment result validity, game-based assessment faces the same issues as traditional psychological tests. Some researchers note that game-based assessment results do not fully equate to test-takers' actual levels of measured traits. Even when game tasks reflect key elements of the measured trait, how test-takers role-play and behave in the game represents only an approximation of their real-life performance (Stănescu et al., 2020). Moreover, environmental elements such as interface colors, character designs, and sound effects, as well as individual factors like prior gaming experience, can all influence assessment results.

In summary, while game-based assessment methods have certain shortcomings, their unique advantages are undeniable. Using games as assessment tools is an increasingly important approach with growing value.

3.1 Evidence-Centered Design

Establishing scientifically valid assessment tools is a prerequisite and foundation for measuring individual psychological characteristics. Therefore, in game-based assessment research, the development and validation of assessment tools have been a focal point for scholars. Evidence-centered design provides the theoretical foundation and has further formed a paradigm for establishing assessment tools.

Mislevy et al. first proposed the conceptual assessment framework in 2003 for the field of educational evaluation—a general model for building assessments consisting of five components: student model, evidence model, task model, assembly model, and presentation model (Mislevy, Almond, & Lukas, 2003; Mislevy, Steinberg, & Almond, 2003), along with four implementation processes: presentation process, response process, scoring process, and task selection process. The conceptual assessment framework and the four processes are collectively known as evidence-centered design (ECD), a broader measurement model to support modern educational assessment. Evidence-centered design is also applicable to developing game-based assessment tools, and Shute (2011) summarized it into three core components: competency model, task model, and evidence model.

3.2 Development of Assessment Tools

First, define the structure of the trait being measured, which involves establishing the competency model. Constructing the competency model requires researchers to identify target traits based on research questions—that is, the knowledge, skills, abilities, or attitudes intended to be measured—and to define

the attributes and characteristics of target traits according to existing theoretical frameworks. This model can be simple, examining a single trait through task completion, or complex, comprehensively assessing several traits within a single game (Sun, 2011).

Second, determine indicators that reflect target traits and scoring rules, which involves establishing the evidence model. The evidence model serves as a bridge between the competency model and task model, aggregating observable values and building predictive models to infer target traits. This is the most crucial component of the ECD framework and can be divided into two parts: statistical rules and statistical models (Shute, 2011). The construction of statistical rules involves selecting indicators from the game that connect to the competency model and establishing scoring rules such as scores or score ratios for test-takers' game performance to obtain observable and quantifiable results. Due to the diversity of games, different games have different data indicators, and even the same game may have different indicators depending on the competency model (Wen, Fu, & Huang, 2019). Typically, indicator selection and establishment rely primarily on relevant research foundations and researchers' experience and expertise without universal standards. Common data indicators include task completion time, number of levels completed, and accuracy rates. Additionally, Nebel and Ninaus (2019) proposed that physiological data could provide deeper insights into players' emotional and cognitive states, suggesting that future research could consider collecting relevant physiological data as assessment indicators.

The construction of statistical models involves defining relationships between observable indicators and the competency model (Feng, 2012). These relationships may be logical or probabilistic. On one hand, researchers can aggregate selected indicator scores based on simple computational rules to directly represent target trait levels. On the other hand, researchers can employ algorithms such as Bayesian networks or random forests to build mathematical models that predict target trait levels based on selected indicators. The choice of statistical model is associated with factors including the target trait, game tasks, and number of indicators.

Generally, logical models are relatively simple. Vendlinski and Stevens (2002) designed a game to assess high school students' chemistry knowledge, requiring test-takers to identify specified chemicals across 23 different scenarios. Test-takers could use experiments, book references, and other methods to assist their judgment within the game, with each scenario scoring 1 point—higher scores indicating higher chemistry knowledge levels. DeRosier et al. (2012) assigned scores to children' s behavioral choices in virtual social situations and calculated total scores to assess children' s social-emotional competence.

Probabilistic models are more complex. In Shute et al.' s (2016) study using Plants vs. Zombies (Use Your Brain) to predict problem-solving skills, they divided the proportion of a certain behavior within its category into levels as a data indicator and established probabilistic relationships between levels and

trait levels. For instance, when a test-taker performed well on a data indicator, the probability of performing well on the target trait was 0.5. Combining performance across numerous data indicators allowed prediction of individual problem-solving ability levels using probabilistic formulas. Currently, no research conclusions indicate that statistical model type significantly affects assessment results, allowing researchers to select appropriate models based on research purposes.

Third, design tasks or situations from which to obtain indicators, which involves establishing the task model. In game-based assessment, the game itself is the assessment task, with the primary purpose of eliciting evidence of test-takers' abilities. This requires defining presentation methods, game task characteristics, difficulty and quantity of game tasks, feasible strategies for completing tasks, and target performance levels (Rupp et al., 2010). Some researchers note that different game types involve different skills and can categorize video games into strategy, adventure, role-playing, action, simulation, and other types (Dickey, 2006). Strategy games involve cognitive abilities, decision-making, and strategic thinking; simulation games relate to problem-solving, self-awareness, and perspective-taking; and role-playing games require imagination, cooperation, and planning (DeRosier & Thomas, 2018b). When selecting games as measurement tools, the first consideration should be the match between target traits and game functions. During research, researchers can either extract predictive indicators from existing games for assessment or design new games based on research purposes.

Using evidence-centered design to establish assessment tools for target traits provides necessary conditions for subsequent data collection, processing, and reliability and validity testing (see Figure 1 [Figure 1: see original paper]).

3.3 Reliability and Validity Testing

As a relatively novel measurement technology, reliability and validity testing is particularly important for game-based assessment, though current research remains limited. In existing studies, validation methods for game-based assessment tools are similar to those for traditional psychological tests, specifically divided into reliability testing and validity testing.

Reliability refers to consistency. High reliability coefficients indicate that assessment tool results are more consistent, stable, and dependable. Typically, internal consistency reliability is tested by calculating Cronbach' s alpha coefficients and intraclass correlation coefficients (Shute & Moore, 2017).

For example, Kim et al. (2016) designed "Physics Playground," a game using 74 levels to assess players' understanding of Newton' s three laws of motion. Players controlled tools such as levers and inclined planes on-screen to move a ball to a target location. If players exhibited behaviors like using ramps, levers, pendulums, or springboards to complete the movement, they were considered to have performed excellently. Researchers calculated the internal correlation of

excellent performance data (r = 0.85) and selected 29 levels with high completion rates for alpha coefficient testing ($\alpha = 0.87$). Additionally, they conducted confirmatory factor analysis on four structural dimensions of excellent performance data, finding small measurement errors for individual dimensions and high internal consistency reliability. These results demonstrate that Physics Playground has desirable reliability.

Validity refers to the extent to which an assessment tool accurately measures its intended trait. Typically, convergent and discriminant validity are tested using external measurement tools (Rupp et al., 2010). Additionally, criterion-related validity is also a focus, reflecting the effectiveness of games in predicting individual behavioral performance in specific contexts.

For instance, Weiner (2019) designed a VR game where test-takers used a head-mounted controller and two hand-held controllers to complete three VR games measuring cognitive abilities including visual speed and accuracy, spatial imagination, and visual tracking. After testing, test-takers completed subtests from the Employee Aptitude Survey (EAS) measuring these three abilities and a Big Five personality test, and researchers also obtained test-takers' academic performance (GPA). Correlations and regression equations were established between VR test scores and external test scores and academic performance. Results showed pairwise correlations among these measures, and VR test scores provided meaningful predictions of academic performance. Results for convergent validity, discriminant validity, and criterion-related validity demonstrate good validity for the VR game.

4 Assessment Practice

Guided by assessment paradigms, researchers have developed various game-based assessment tools to evaluate individual abilities and behaviors. Cognitive and non-cognitive abilities, as two sides of the same coin, are crucial for personal development (Li & Zhao, 2017) and thus frequently become targets of psychological testing. Currently, game-based assessment is widely applied in evaluating individual cognitive abilities (De Klerk et al., 2015) and also offers unique advantages in predicting non-cognitive abilities.

4.1 Cognitive Ability Assessment

Cognitive abilities are those required when individuals reconstruct and apply knowledge, involving basic cognitive functions such as perception, memory, and attention, as well as higher-order functions like reasoning, judgment, imagination, and problem-solving. Game-based assessment provides new approaches for evaluating cognitive abilities, with applications in both cognitive ability evaluation and diagnosis.

Sun et al. (2018) used the Sokoban game to predict individuals' reasoning ability and mathematics achievement, extracting 23 features including first-step

time ratio, box completion ratio, thinking step ratio, repeated step ratio, and steps deviating from optimal path to build a random forest model. They validated the model' s predictive effectiveness by calculating precision, recall, and other metrics. Shute et al. (2016) used Plants vs. Zombies (Use Your Brain) to predict problem-solving skills, including four dimensions: analyzing conditions and constraints, developing solutions, effectively using resources and tools, and monitoring and adjusting processes. Based on test-takers' operations to fend off zombies—for example, "using plant food when there are more than five zombies" was considered effective resource utilization—they transformed target behaviors/total behaviors into frequencies, extracted 32 features, and built a Bayesian network model. The prediction results showed significant correlations with Raven' s Progressive Matrices and a simulated basketball task. Argumentative reasoning ability is another important cognitive skill. Researchers used the Seaball—Semester at Sea game, requiring children to answer whether presented foods were junk food and select reasons from multiple options, then scored 48 items to obtain a total game score (Song & Sparks, 2019). Higher scores indicated stronger argumentative reasoning ability. Students' scores on the game assessment showed moderate correlations with the Cognitively Based Assessment of, for, and as Learning (CBAL) test scores, demonstrating the game' s discriminant and convergent validity. Scores also significantly correlated with teacher-reported student achievement and ratings of argumentative writing ability, demonstrating criterion-related validity.

Beyond assessing cognitive abilities in general populations, game-based assessment has also been used for cognitive diagnosis in populations with cognitive impairments. Manera et al. (2015) used a "Kitchen and Cooking" game task to assess elderly individuals with mild cognitive impairment and Alzheimer' s disease, requiring test-takers to click on screens to prepare dishes. The game consisted of three processes: identifying raw materials, planning preparation procedures, and actual operation, involving perceptual, planning, and practical abilities. Final judgment indicators included completion time and error frequency, with game performance significantly correlating with results from tests of overall cognitive function, attention and thinking, executive function, and memory, thereby validating the game' s validity. Flynn et al. (2019) administered and monitored a cognitive impairment summer camp for children using a set of game tasks comprising a repeatable cognitive detection tool, including perceptual discrimination tasks (clicking correct targets on screen) and navigation tasks (tilting an iPad to guide an avatar around obstacles). Using adaptive algorithm tools to automatically record individual performance on single and multiple tasks, they collected 20 response indicators. Through longitudinal data results, more comprehensive and accurate assessments can be conducted during treatment processes for improving cognitive-neurological disorders.

4.2 Non-Cognitive Ability Assessment

Due to the difficulty of measuring non-cognitive abilities that reflect social characteristics and personality traits, attention to these abilities has lagged behind. However, as the importance of non-cognitive abilities in individual development has become increasingly apparent, the role of game-based assessment in evaluating non-cognitive abilities has recently gained attention.

Researchers used the "Zoo U" game to assess children' s social-emotional competence. During gameplay, children interacted with virtual characters in a school-like story world to complete situational choice problems presented across six virtual social scenarios, thereby evaluating six abilities: communication, cooperation, empathy, emotion regulation, impulse control, and social activities. Results showed positive correlations among internal consistency across the six dimensions. Additionally, researchers obtained teacher-reported social skills and behavior rating scale scores, disciplinary actions, and academic adjustment data for tested students. Children with lower game scores exhibited more social, behavioral, and academic problems (DeRosier et al., 2012; DeRosier & Thomas, 2018a). Furthermore, assessments of individual teamwork competency (Guenaga et al., 2015) and personal cooperative behavior (Keil et al., 2017) can also be implemented through games.

Personality trait measurement has also received scholarly attention. Nimwegen et al. (2011) collaborated with a game studio and a human resources consulting company to develop a game for assessing individual compliance, where test-takers simulated making decisions and expressing opinions in a corporate environment. The actions chosen in the story actually represented scores on a four-point Likert scale. The Poptropica island quest game can predict individual persistence traits (DiCerbo, 2014). Researchers selected island levels with first-pass rates below ten percent as difficult levels, using time spent on task events and number of task completions as persistence indicators. They built a confirmatory factor analysis model using two assessment indicators from three island quests, obtaining good fit indices. The well-established ultimatum game paradigm in psychology can assess individual altruism traits, while the dictator game can measure fairness traits (Baumert et al., 2014). Based on behavioral performance in games, researchers can also assess individual proactive and reactive aggression traits (McCreery et al., 2019).

However, some scholars have reached different conclusions in the personality trait domain. Dalveren et al. (2015) used a surgical ward navigation game to test personality traits, requiring test-takers to reach ten different target locations based on a map. Referring to Jung' s personality types (Myers-Briggs Type Indicator, MBTI), they selected indicators such as reaction time, distance traveled, number of wrong paths taken, number of wall collisions, and task success rate as game behavior metrics. However, analyzing correlations between these personality types and players' in-game performance revealed no significant relationships between game parameters and players' personality types, challeng-

ing the effectiveness of game-based assessment in predicting personality traits.

5 Future Research Directions

In summary, this paper has systematically reviewed the concepts, paradigms, and practical progress of game-based assessment. This synthesis is not only valuable for understanding this emerging technology but also provides important guidance for subsequent research. However, research on game-based assessment remains in its initial stages. Future scholars can further enrich related research from three perspectives: assessment task design, data analysis, and practical application.

5.1 Assessment Task Design

Early studies primarily utilized existing commercial games such as Plants vs. Zombies and Sokoban (Shute et al., 2016; Sun et al., 2018) to explore associations between performance on these games and certain psychological traits. Currently, an increasing number of researchers are attempting to develop topic-specific games by integrating game mechanics, content, and assessment based on the evidence-centered design framework (e.g., Song et al., 2020). This approach is more targeted for assessing corresponding psychological traits and yields more accurate and complete content. The emergence of some universal game frameworks, such as Minecraft, has also lowered the barrier to game development to some extent.

However, most current game-based assessments still employ linear design patterns, presenting identical game scenarios and content to different test-takers, which results in time-consuming assessments with relatively monotonous content. Consequently, some researchers have proposed non-linear game models. One manifestation is branching design, where different behaviors lead to different game scenarios. Bacos et al. (2018) used an interactive narrative game with branching storylines to measure individuals' counterfactual thinking, suggesting that branching design may provide direction for assessing more advanced traits. Another manifestation is adaptive design, which selects game levels of appropriate difficulty based on estimates of test-takers' abilities, substantially improving assessment efficiency. Wilson et al. (2006) developed an adaptive game software called The Number Race to remediate dyscalculia in children, assessing children' s baseline calculation abilities and presenting problems suited to their performance levels to train calculation skills. Although this game' s primary purpose was training rather than assessment, it offers important insights. Since few researchers in the game-based assessment field have conducted studies on adaptive game design, future researchers could explore this further by drawing on related research from game-based training and game-based learning domains.

Additionally, multi-player large-scale game design provides direction for simultaneously measuring multiple individuals and multiple traits. Annetta et al. (2010)

developed a multiplayer educational gaming application (MEGA) to assess 21st-century digital literacy, creative thinking, executive function, and communication skills in students, evaluating four factors: teacher-student interaction, peer discussion, engagement levels, and time spent playing. Early studies featured relatively simple evaluations of game behavior and game scenario design, but advances in computer technology will bring more possibilities for designing non-linear and multi-player multi-trait game models.

5.2 Assessment Data Analysis

Early studies primarily used outcome data, but attention to process data has gradually increased, with a trend toward integrated application of process and outcome data. De Klerk et al. (2015) summarized 31 research findings, among which ten studies used game outcome data, six used process data, and the remainder used both, indicating that researchers' utilization of process data remains limited. With advances in computer technology, the significant advantages of machine learning in data processing have become increasingly apparent, especially since game-based approaches generate large volumes of data that traditional statistical methods cannot maximize (Csapó et al., 2012). Machine learning algorithms can help researchers build more complex models during the outcome evaluation stage.

Many researchers have already introduced algorithms such as Bayesian networks, decision trees, and random forests to build predictive models. Sun et al. (2018) selected test-taker samples from the top 25% and bottom 25% of test scores for feature extraction and model building, extracting 23 feature indicators from Sokoban game performance as feature values for classification datasets. After randomly dividing training and test sets, they trained and classified the dataset to establish relationships between Sokoban performance and reasoning ability and mathematics achievement. Future research could also consider combining convolutional neural networks to process image data, allowing game data to provide more information, and adopting unsupervised learning types in machine learning to explore inherent grouping patterns or rules within data, enriching analysis methods for assessment results. It should be noted that although this processing approach has good statistical significance, machine learning is a data-driven modeling process aimed at maximizing predictive accuracy and sometimes cannot simultaneously consider the meaning and structure of features themselves (Mayer et al., 2014). Conclusions derived solely from data-driven approaches may lack practical significance (Wu, Hu, & Zhao, 2015). How to combine theoretical foundations with machine learning methods requires more in-depth analysis and research.

5.3 Practical Application of Assessment

Regarding assessment content, early game-based assessment primarily focused on evaluating individuals' mastery of knowledge and skills. Unlike paper-and-

pencil tests, researchers integrated assessment points into games, enabling test-takers to demonstrate their understanding and application of knowledge and skills during gameplay. Game-based assessment is particularly effective when evaluating knowledge and skills that require deeper understanding and application, such as mathematics, physics, medical first aid, and architectural design (De Klerk et al., 2015). As game-based assessment has developed, its role in assessing cognitive and non-cognitive abilities has also gained widespread attention. Game-based assessment can avoid certain shortcomings of traditional psychological tests, such as susceptibility to social desirability effects and lack of process data, making it increasingly favored by researchers. More researchers are using games to assess and study individual psychological traits, and many companies have designed their own games for talent recruitment to evaluate candidates' abilities and personalities. However, selecting data indicators from game tasks to represent individual personality characteristics remains highly challenging, resulting in relatively simple and limited current practice in non-cognitive ability assessment. Future researchers should explore this direction more deeply.

Regarding application contexts, recent research attempts using game-based assessment in clinical evaluation and treatment have brought new value to this technology. Hautala et al. (2020) developed a set of online game tasks to assess and screen for reading disabilities in lower-grade students. Song et al. (2020) designed a mobile game called "CoCon" to assess cognitive functions in children and adolescents and planned to extend CoCon' s use to screening clinical populations with severe cognitive control problems. Game-based assessment holds high application value in subsequent treatment interventions because it can establish automatic scoring systems, meticulously record changes in individual levels during interventions, and automatically adjust task difficulty through adaptive algorithms. In June 2020, the U.S. Food and Drug Administration approved a game called EndeavorRx as a prescription treatment for children with ADHD, reflecting the enormous practical potential in this field. How to combine theory with practice to maximize the functionality of game-based assessment requires researchers' continuous efforts.

In practical applications, researchers are also increasingly focusing on detailed issues such as assessment indicator selection. In DiCerbo' s (2014) study predicting individual persistence through the Poptropica island quest game, four behaviors were pre-selected as assessment indicators, but small-sample testing revealed unstable relationships and low validity for these four indicators. Consequently, the final assessment indicators selected were time spent on task events and number of task completions. Overly simplistic use of two indicators to represent individual persistence raises questions about the game' s assessment validity. Therefore, when designing evidence models, it is necessary to predetermine an appropriate number of behaviors to include as assessment indicators and carefully define the connections between behaviors, evidence, and constructs. Some researchers note that behavioral indicators sometimes cannot be determined during initial evidence model design (DiCerbo, 2017) and

can be identified through iterative approaches, developing and confirming hypotheses about which elements in log files constitute evidence from test-takers' task completion processes. Other researchers have focused on default assumptions in previous game-based assessment research, such as games' promotion of assessment motivation and engagement, and their impact on test anxiety and academic performance (Verma et al., 2019).

Furthermore, assessment result errors caused by individual differences among test-takers have also attracted researchers' attention. Compared to females, males play various types of games more frequently and persistently, are more familiar with common game patterns and rules, which may help them perform better in games. Researchers have measured different individual differences such as gender, enjoyment, gaming self-efficacy, and gaming time to examine their impact on assessment scores, but have not reached unified conclusions (Sanchez & Langer, 2020). Kim and Shute (2015) compared results between males and females and between experienced and inexperienced gamers, finding that experienced gamers had greater advantages on some indicators, males had higher game completion rates, and significant gender differences existed on two key indicators. These advantages may manifest in only a few game indicators and can be identified and adjusted through testing. Game-based assessment tools must minimize the impact of individual differences among players to accurately measure test-takers' abilities, and sufficient learning opportunities should be provided when necessary to reduce differences arising from gaming backgrounds. Oranje et al. (2019) also caution that cultural contexts of games and players need attention in game-based assessment. When applying a game designed and validated in one culture to groups in another culture, evaluation methods similar to traditional assessment should be adopted to ensure cross-cultural measurement invariance.

Although game-based psychological assessment is still in its infancy and domestic research remains scarce, its tremendous potential in the field of psychological measurement is foreseeable. The research paradigm centered on evidence-centered design provides guidance for establishing assessment tools, and research practice in cognitive and non-cognitive abilities based on this paradigm has validated the effectiveness of game-based assessment technology. With continuous advances in computer and gaming technology, game-based assessment is expected to play important roles in multiple fields including educational evaluation, psychological measurement, and human resource management in the future.

References

冯翠典. (2012). “以证据为中心”的教育评价设计模式简介. 上海教育科研, 8, 12–16.

李丽, 赵文龙. (2017). 家庭背景、文化资本对认知能力和非认知能力的影响研究. 东岳论丛,38(04) ,142–150.

孙海洋. (2011). 以证据为中心的设计模式及其对口语考试设计的启示. 中国考试, 7, 20–26.

孙鑫, 黎坚, 符植煜. (2018). 利用游戏 log-file 预测学生推理能力和数学成绩——机器学习的应用. 心理学报, 50(7), 761-770.

温迎, 付玉, 黄贝萱. (2019). 电子游戏测评综述. 电脑知识与技术, 15(8), 184-186.

吴忭, 胡艺龄, 赵玥颖. (2015). 如何使用数据：回归基于理解的深度学习和测评——访国际知名学习科学专家戴维·谢弗. 开放教育研究, 25(1), 6-14.

吴宇. (2015). 浅析严肃游戏在中国的发展. 科技视界, 10, 135+227.

杨振芳, 孙贻文. (2015). 游戏化招聘：人才选拔的新途径. 中国人力资源开发, 342(24), 47-52

Annetta, L., Cheng, M.T., & Holmes, S. (2010). Assessing twenty-first-century skills through a teacher-created video game for high school biology students. *Research in Science & Technological Education*, 28, 101-

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance? *Computers & Education*, 83, 57-63.

Bacos, C. A., McCreery, M. P., & Laferriere, J. R. (2018). Interactive narratives, counterfactual thinking, and personality in video games. *HCI International 2018 -Posters' Extended Abstracts*, 850, 340-347.

Baron, T. (2017). *An architecture for designing content agnostic game mechanics for educational burst games* (Unpublished doctoral dissertation). Arizona State University.

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment*, 30(3), 178-192.

Caballero-Hernández, J. A., Palomo-Duarte, M., & Dodero, J. M. (2017). Skill assessment in learning experiences based on serious games: A systematic mapping study. *Computers & Education*, 113, 42-60.

Csapó B., Ainley J., Bennett R.E., Latour T., & Law N. (2012) Technological issues for computer-based assessment. In: P. Griffin, B. McGaw, & E. Care (eds), *Assessment and Teaching of 21st Century Skills*(pp. 143-230). Dordrecht: Springer

Dalveren, G.G.M, Cagiltay, N., & Ozcelik, E. (2015). Personality type indicator models in serious games: A case study in a surgical navigation game. *Information Technology Based Higher Education and Training*, 1-4

De Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education*, 85, 23-34.

DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*, 1, 1-7.

DeRosier, M. E., & Thomas, J. M. (2018a). Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Developmental Psychology*, 55, 52–61.

DeRosier M. E., & Thomas, J. M. (2018b). Video games and their impact on teens' mental health. In: M. Moreno, & A. Radovic (eds). *Technology and Adolescent Mental Health*(pp. 237–253). Cham, Switzerland: Springer

DeRosier, M. E., & Thomas, J. M. (2019). Hall of heroes: A digital game for social skills training with young adolescents. *International Journal of Computer Games Technology*, 6, 1–12.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society*, 17 (1), 17–28.

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology*, 18, 7–18.

Dickey, M. D. (2006). Game design and learning: A conjectural analysis of how massively multiple online role- playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development*, 55(3), 253–273.

Flynn, R. M., Colón-Acosta, N., Zhou, J., & Bower, J. (2019). A game-based repeated assessment for cognitive monitoring: Initial usability and adherence study in a summer camp setting. *Journal of Autism and Developmental Disorders*, 49(5), 2003–2014.

Gamberini, L., Marchetti, F., Martino, F., & Spagnolli, A. (2009). Designing a serious game for young users: The case of happy farm. *Studies in Health Technology and Informatics*, 144, 77-81.

Gee, J. P. (2008). Learning and games. In: K. Salen(ed), *Good Video Games and Good Learning*(pp. 21–40). New York: Peter Lang Publishing

Guenaga, M., Eguíluz, A., & Rayon, A. (2015). A serious game to develop and assess teamwork competency. *International Symposium on Computers in Education, SIIE 2014*, 183–188.

Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V., Latvala, J. M., & Richardson, U. (2020). Identification of reading difficulties by a digital game-based assessment technology. *Journal of Educational Computing Research*, 58(5), 1003–1028.

Heinzen T.E., Landrum R.E., Gurung R.A.R.,& Dunn D.S. (2015). Game-based assessment: The mash-up we' ve been waiting for. In: T. Reiners, & L. Wood(eds), *Gamification in Education and Business*(pp. 201–207). Cham, Switzerland: Springer

Keil, J., Michel, A., Sticca, F., Leipold, K., Klein, A. M., Sierau, S., . . . White, L. O. (2017). The Pizzagame: A virtual public goods game to assess cooperative behavior in children and adolescents. *Behavior Research Methods*, 49(4), 1432–1443.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game- based assessments in physics playground. *International Journal of Testing*, 16(2), 142–163.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, 87, 340–356.

Kim, Y. J.,& Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In: D. Ifenthaler, & Y. J. Kim(eds), *Game-Based Assessment Revisited. Advances in Game-Based Learning*(pp. 3–11). Cham, Switzerland: Springer

Manera, V., Petit, P. D., Derreumaux, A., Orvieto, I., Romagnoli, M., Lyttle, G., . . . Robert, P. H. (2015). "Kitchen and cooking" , a serious game for mild cognitive impairment and Alzheimer' s disease: A pilot study. *Frontiers in Aging Neuroscience*, 7–24.

Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, 33(2), 137–150.

Mayer I., Dierendonck D.V, Ruijven T.V, & Wenzler I. (2014). Stealth assessment of teams in a digital game environment. In: A. De Gloria(ed), *Lecture Notes in Computer Science*(Vol. 8605, pp. 224–235). Cham, Switzerland: Springer

McCreery, M., Krach, S., Bacos, C., Laferriere, J., & Head, D. (2019). Can video games be used as a stealth assessment of aggression?: A criterion-related validity study. *International Journal of Gaming and Computer-Mediated Simulations*, 11, 40–49.

Mislevy, R., Almond, R., & Lukas, J. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 1, 1–29.

Mislevy, R., Steinberg, L., & Almond, R. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, 1(1), 3–62.

Mislevy, R. J., Behrens, J., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*, 4, 11–

Nebel, S., & Ninaus, M. (2019). New perspectives on game-gased assessment with process data and physiological signals. In D. Ifenthaler. & Y. J. Kim. (Eds.), *Game-Based Assessment Revisited* (pp. 141–161). Cham, Switzerland: Springer

Nicholson S. (2015). A recipe for meaningful gamification. In: T. Reiners, & L. Wood(eds), *Gamification in Education and Business*(pp. 1–20). Cham, Switzerland: Springer

Nimwegen, C., Oostendorp, H., Modderman, J., & Bas, M. (2011). A test case for GameDNA: Conceptualizing a serious game to measure personality traits. *16th International Conference on Computer Games*, 217–222

Oranje, A., Mislevy, B., Bauer, M. I., & Jackson, G. T. (2019). Summative game-based assessment. In: D. Ifenthaler, & Y. J. Kim(eds), *Game-Based Assessment Revisited* (pp. 37–65). Cham, Switzerland: Springer

Rupp, A., Gushta, M., Mislevy, R., & Shaffer, D. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, 8(4), 1–48.

Sanchez, D. R., & Langer, M. (2020). Video game pursuit (VGPu) scale development: Designing and validating a scale with implications for game-based learning and assessment. *Simulation & Gaming*, 51(1), 55–86.

Schoedel, R., Au, Q., Völkel, S., Lehmann, F., Becker, D., Buehner, M., . . . Stachl, C. (2018). Digital footprints of sensation seeking: A traditional concept in the big data era. *Zeitschrift für Psychologie*, 226(4), 232–245.

Shute, V. J , & Moore,G.R.(2017). Consistency and validity in game-based stealth assessment. In J. Hong, & W. L. Robert (Eds.), *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*(pp. 31–51). Charlotte: Information Age

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher(Eds.), *Computer Games and Instruction*(pp. 503–523). Charlotte: Information Age

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63, 106–117.

Song, H., Yi. D. J.,& Park, H. J. (2020). Validation of a mobile game-based assessment of cognitive control among children and adolescents. *PLoS ONE*, 15(3), 1–18.

Song, Y., & Sparks, J. R. (2019). Measuring argumentation skills through a game-enhanced scenario-based assessment. *Journal of Educational Computing Research*, 56(8), 1324–1344.

Stănescu, D. F., Ioniță, C., & Ioniță, A. M. (2020). Game-thinking in personnel recruitment and selection: Advantages and disadvantages. *Postmodern Openings / Deschideri Postmoderne*, 11(2), 267–276.

Turan, Z., & Meral, E. (2018). Game-based versus to non-game-based: The impact of student response systems on students' achievements, engagements and test anxieties. *Informatics in Education*, 17(1), 105-116.

Vendlinksi, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *The Journal of Technology, Learning and Assessment*, 1(3).

Verma, V., Baron, T., Bansal, A., & Amresh, A. (2019). Emerging practices in game-based assessment. In: D. Ifenthaler, & Y. J. Kim(eds), *Game-Based Assessment Revisited* (pp. 327–346). Cham, Switzerland: Springer

Weiner, E. J. (2019). Cognitive abilitiy in virtual reality: Validity evidence for VR game-based assessment. (Unpublished master's thesis). San Francisco State University.

Wilson, A. J., Dehaene, S., Pinel, P., Revkin, S., Cohen, L., & Cohen, D. (2006). Principles underlying the design of "The Number Race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, 2(1), 2-19.

Note: Figure translations are in progress. See original paper for figures.