ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Ohlms, Marie L.; Melchers, Klaus G.; Kanning, Uwe P.

Article — Published Version

Can we playfully measure cognitive ability? Construct-related validity and applicant reactions

International Journal of Selection and Assessment

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Ohlms, Marie L.; Melchers, Klaus G.; Kanning, Uwe P. (2023) : Can we playfully measure cognitive ability? Construct-related validity and applicant reactions, International Journal of Selection and Assessment, ISSN 1468-2389, Wiley, Hoboken, NJ, Vol. 32, Iss. 1, pp. 91-107, https://doi.org/10.1111/ijsa.12450

This Version is available at:
https://hdl.handle.net/10419/290354

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

http://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Received: 21 November 2022 | Revised: 8 August 2023 | Accepted: 10 August 2023

DOI: 10.1111/ijsa.12450

RESEARCH ARTICLE

International Journal of Selection and Assessment | WILEY

Can we playfully measure cognitive ability? Construct-related validity and applicant reactions

Marie L. Ohlms[1] | Klaus G. Melchers[1] | Uwe P. Kanning[2]

[1]Institute of Psychology and Education, Ulm University, Ulm, Germany

[2]Faculty of Economics and Social Sciences, University of Applied Sciences Osnabrück, Osnabrück, Germany

Correspondence
Marie L. Ohlms, Institute of Psychology and Education, Ulm University, Albert-Einstein-Allee 41, D-89069 Ulm, Germany.
Email: marie.ohlms@uni-ulm.de

Funding information
Studienstiftung des Deutschen Volkes

Abstract

We developed a game-based assessment (GBA) measuring cognitive ability for use in personnel selection and examined its construct-related validity. Moreover, applicant reactions toward this GBA were compared with a paper-pencil-based ability test. Both assessment tools were designed to measure verbal, numerical, and figural ability. $N = 183$ participants completed the GBA, the paper-pencil test, and questions capturing applicant reactions and personality. We found a strong positive correlation of 0.51 between the overall GBA and paper-pencil test scores, showing evidence for its construct-related validity. Applicant reactions toward the GBA were consistently worse compared with the paper-pencil test. Furthermore, males and individuals with more video game experience held more positive perceptions than females and individuals with less video game experience.

KEYWORDS

applicant reactions, cognitive ability, game-based assessment, gamification, personnel selection

Practitioner points

- Game-based assessment (GBA) is an innovative field of personnel selection that is increasingly used in practice, although empirical evidence on psychometric properties of GBAs and on applicant reactions is still scarce to date.
- The present study found a strong correlation between the GBA and a traditional intelligence test, indicating that a GBA can provide a valid measurement of cognitive ability in the context of personnel selection.
- Applicant reactions toward the GBA were consistently worse compared with a paper-pencil test measuring the same cognitive abilities.
- Applicant reactions differed between males and females, suggesting that further research is needed to investigate relevant design elements of GBAs affecting their psychometric properties and applicant reactions.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *International Journal of Selection and Assessment* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Increasing digitalization in selection and assessment has led to far-reaching changes in the personnel selection processes of organizations in recent years (Ryan & Derous, 2019). In line with this, an innovative field that has recently received considerable attention in the domain of personnel selection is *game-based assessment* (GBA; Landers & Sanchez, 2022).

It has been argued that GBAs offer manifold applications in the selection process, as GBAs may serve to measure a variety of different constructs depending on their content and specific design, and that even the simultaneous measurement of multiple relevant characteristics within one GBA may be possible (Bhatia & Ryan, 2018). However, GBAs in general can take on very different appearances and may differ considerably in their content and design (Bhatia & Ryan, 2018). For instance, GBAs can vary in length, ranging from a few minutes to several hours. In addition, a variety of game genres (e.g., action, adventure, strategy) and scoring methods (i.e., path- or outcome-based) can be used. Different multimedia styles are also conceivable (Fetzer et al., 2017). Accordingly, GBAs cannot be perceived as a single entity, which means evidence concerning the validity of one GBA cannot be generalized to other GBAs.

In recent years, GBAs have become increasingly popular in applied settings. This may especially be due to the numerous advantages postulated by proponents of this game-based method. For example, it is often claimed that GBAs should positively influence applicant reactions as well as enjoyment during test taking and thereby improve the recruiting function of selection instruments compared with more traditional methods (Bhatia & Ryan, 2018; Weidner & Short, 2019). Particularly, considering the war for talent (Chambers et al., 1998), applicant reactions are becoming increasingly important when it comes to attracting top talent to work for one's organization. Thus, the hoped-for positive impact of GBAs on applicant reactions might help organization to succeed in the competition for top talent. Another potential advantage of GBAs that is often mentioned in literature, is that they may be less susceptible to socially desirable response behavior, faking, and biasing influences of test anxiety on test performance, as the actual purpose of the assessment should be less obvious (Bhatia & Ryan, 2018; Weidner & Short, 2019).

Despite the increasing use of GBAs in practice, empirical evidence concerning their validity as well as support for their suggested advantages is still scarce to date, as was revealed in a recent review of the relevant literature (Ramos-Villagrasa et al., 2022). For example, only limited research exists on the validity of GBAs measuring cognitive abilities, even though many GBAs are designed to measure such abilities (Woods et al., 2020; but see Landers et al., 2022, for a valuable exception). Therefore, further research is needed to investigate the psychometric properties of GBAs. The accumulation of evidence from different GBAs might also allow us to understand whether different game genres or certain design elements influence the validity of GBAs. This is important because, given the large differences between different GBAs, it is unclear to which degree results of one GBA can be generalized to another.

Thus, as the empirical knowledge about GBAs (as described in detail below) is limited and given that results from one GBA cannot simply be generalized to another GBA, the first goal of the present study was to develop and construct validate a GBA from the adventure genre targeting cognitive ability for use in personnel selection. Specifically, the validity of the GBA was investigated by considering its correlation with a traditional paper-pencil cognitive ability test which aims to measure the same abilities. We focused on cognitive ability because especially the use of GBAs that measure cognitive ability offers promising opportunities as cognitive ability is among the best predictors of job performance (Sackett et al., 2022; Schmidt & Hunter, 1998). In addition, traditional ability tests are less accepted by applicants compared with other personnel selection tools (Anderson et al., 2010). Therefore, the second goal was to evaluate whether applicant reactions toward the cognitive GBA were more positive than those to a traditional paper-pencil test measuring the same abilities. Furthermore, the measurement of cognitive ability may be impaired, for example, by applicants' test anxiety, which introduces measurement error (McCarthy & Goffin, 2005). Therefore, we also compared test takers' anxiety in the GBA versus the paper-pencil test. Taken together, the results of the current study aim to further researchers' and organizations' understanding of potential advantages and disadvantages of GBAs.

2 | PREVIOUS RESEARCH ON GAME-RELATED ASSESSMENTS REGARDING VALIDITY AND APPLICANT REACTIONS

Game-related assessments (GRAs), which incorporate GBA, gamification, and gamefully designed assessment (GDA) have already received some attention in selection and assessment research for a longer time (e.g., Kleinmann & Strauß, 1998), but the steady increase in associated empirical research mainly started in the last decade, leading to several studies in this area (see the review by Ramos-Villagrasa et al., 2022). GBA refers to games that are used as stand-alone instruments to measure the constructs of interest (Bhatia & Ryan, 2018; Chamorro-Premuzic et al., 2016; Fetzer et al., 2017). In contrast, gamification describes the introduction of game elements into existing diagnostic instruments, and gameful design refers to the use of game elements to develop an entirely new selection instrument (Landers & Sanchez, 2022). Therefore, gamification and GDA can also be used in procedures that are not games (Deterding et al., 2011; Landers & Sanchez, 2022). Thus, the crucial distinction between gamification and GDA on the one hand, and GBA on the other hand, is that the former are (re)design strategies, whereas GBA represents an assessment method (Landers & Sanchez, 2022).

As described below, the few available studies on GRAs yielded several promising findings so far (e.g., Georgiou et al., 2019; McChesney, et al., 2022). For instance, Georgiou et al. (2019) found moderate support for the construct-related validity of their gamefully

designed version of a traditional situational judgement test (SJT) called Owiwi that was designed to assess applicants' soft skills. In a follow-up study, they also found several significant correlations between scores from Owiwi and self-reported job and academic performance (Nikolaou et al., 2019). Additionally, the gamefully designed SJT was superior to the text-based SJT version in terms of various applicant reactions, except for job-relatedness (Georgiou & Nikolaou, 2020; Georgiou, 2021; Gkorezis et al., 2020). Choosing a different gameful design approach, Barends et al. (2019) used virtual behavior cues in a GDA to measure Honesty-Humility. They found modest support for the construct-related validity of their GDA with Honesty-Humility scores from a traditional HEXACO personality test. However, in contrast to common expectations, the GDA was not less susceptible to faking.

A few studies investigated the effects of introducing game-elements such as storification (i.e., converting an existing selection instrument into a story; Landers & Collmus, 2022) and game-framing (i.e., framing an assessment test as a game; Collmus & Landers, 2019; McChesney et al., 2022) on the validity and applicant reactions compared with more traditional instruments. However, these studies found only mixed support for the introduction of various game-design elements. Given the mixed results in this area, further research is needed to better understand the impact of different gamification and game-design techniques on psychometric properties and applicant reactions (Landers & Sanchez, 2022).

While—as described above—some studies have already examined the impact of gamification and GBAs on applicant reactions (see Ramos-Villagrasa et al., 2022, for a review), current empirical evidence on the psychometric properties of GBAs (especially GBAs to assess cognitive ability) is far more limited in the personnel selection and assessment context. In one of the first studies to address this issue, Landers et al. (2022) reported evidence to support the construct- and criterion-related validity of a theory-driven GBA called Cognify. This GBA consists of seven separate web-based mini-games designed to assess applicants' cognitive ability. Furthermore, applicant reactions to Cognify were more positive than to a traditional cognitive ability test battery. Nevertheless, as stated repeatedly, further research validating GBAs in general and for cognitive ability in particular is needed for drawing solid conclusions about the psychometric properties of GBAs (see Bhatia & Ryan, 2018; Ramos-Villagrasa et al., 2022; Woods et al., 2020). More specifically, more research is needed on the construct- and criterion-related validity of other GBAs that are from different genres and that use different design elements, as this would allow for the development of a better understanding of aspects that contribute to improve psychometric properties of GBAs (Ramos-Villagrasa et al., 2022). Furthermore, given the large differences between different GBAs, we cannot simply generalize the results of existing studies, such as those from Landers et al. (2022), to other different GBAs. In addition, as pointed out in the review by Ramos-Villagrasa et al. (2022), further research examining GBAs in working samples is needed, as most of the few available studies so far used student samples.

2.1 | Relationship between video game performance and cognitive ability

Though empirical evidence on the use of video games or computer simulations to measure cognitive ability in a personnel selection context is scarce (Woods et al., 2020), there are a number of studies examining the relationship between performance in commercial video games and cognitive ability (see Quiroga & Colom, 2020, for a detailed overview). Several studies found moderate to high correlations between performance in such commercial video games and cognitive ability (e.g., Baniqued et al., 2013; McPherson & Burns, 2007; Quiroga et al., 2019). For instance, Ángeles Quiroga et al. (2015) investigated to what extent so-called brain games (i.e., games that also serve a learning purpose and not just entertainment) are an appropriate measure of cognitive ability. They found a high correlation of $r = .93$ between latent video game performance and intelligence. In a follow-up study, Quiroga et al. (2019) tested whether nonbrain games (i.e., games that serve only entertainment purposes) are also suitable for assessing intelligence. Using a video game battery consisting of 10 games of different genres for Wii-U and iPad, they found that performance in nonbrain games was also associated with intelligence and found a correlation of $r = .79$ between latent video game performance and cognitive ability.

The aforementioned results, as well as initial validation studies of GBAs assessing cognitive ability in the personnel selection context (e.g., Landers et al., 2022), undoubtedly provide impetus to consider video games as a tool for measuring cognitive ability in the context of personnel selection and assessment (Quiroga & Colom, 2020).

3 | DEVELOPMENT OF A GBA TO MEASURE COGNITIVE ABILITY

The development of the GBA used in the present study to measure cognitive ability was based on the Cattell–Horn–Carroll (CHC) model (McGrew, 2009). In particular, we focused on fluid intelligence when designing the various tasks within the GBA, because, among the different second-stratum abilities of the CHC model, fluid intelligence shows the strongest correlations with the overarching general intelligence factor (Wilhelm & Schroeders, 2019). As fluid intelligence is differentiated into three Stratum I abilities—sequential reasoning (verbal), induction (numerical), and inductive reasoning (figural)—the items and game elements were chosen to test abilities in these three domains. Furthermore, the game elements and the specific design of the GBA were selected based on the existing literature on gamification to develop an entertaining and diagnostically feasible instrument (e.g., Bedwell et al., 2012; Fetzer et al., 2017; Landers et al., 2018).

As pointed out by Landers et al. (2018) in their gamification framework, game elements (e.g., storylines or sounds) can influence the desired outcomes of gamification, such as, in our case, a psychometrically valid measurement of cognitive ability. Hence, we paid attention to the theory-driven choice of appropriate game elements. Accordingly, based on aspects suggested by Fetzer et al. (2017), which need to be addressed when developing a GBA for selection purposes, the present GBA was designed considering the following gamification design principles: Target group, length, genre, multimedia style, scoring, linear versus nonlinear gameplay, branding, and candidate feedback. For example, regarding the target group, we tried to ensure that the game design did not unfairly advantage or disadvantage certain applicant groups, as our GBA should be suitable for all potential applicants. Drawing on prior research, the GBA used in the present study was developed to avoid potential subgroup bias by focusing on the following criteria: Moderate complexity, different item groups (i.e., low consistency between different items or modules), and ensuring that there is no need to transfer between tasks, but rather that the individual tasks are independent of one another (Quiroga et al., 2009, 2011).

The GBA used in the current study was developed in cooperation with a company that is specialized in the field of e-sports and gaming. More precisely, the development process started with the implementation of the game environment in Minecraft and the conceptualization and programming of the storyline. According to this storyline, participants were instructed to release a cursed country by finding a crystal hidden at the top of a multilevel tower. After the development environment had been created, the individual problem-solving items were implemented into the game. The different items were developed by subject matter experts (i.e., psychologists) who used established cognitive ability measures as a starting point (e.g., Liepmann et al., 2007, 2012).

In contrast to Landers et al. (2022), who used a GBA from the mini-game genre to measure quantitative knowledge, reading and writing, fluid reasoning, and processing speed, we developed a GBA from the adventure genre, designed to measure verbal, numerical, and figural ability. Accordingly, our GBA differs fundamentally from that of Landers et al. (2022) in terms of the targeted cognitive abilities as well as the design of the GBA.

3.1 | Construct-related validity

To meet the requirements of a construct-valid personnel selection instrument, GBA performance should correlate positively with a traditional intelligence test measure that captures the same Stratum II ability as our GBA (i.e., fluid intelligence). Accordingly, we predict that:

Hypothesis 1: There will be a large positive correlation between GBA performance and cognitive ability measured by a traditional cognitive ability test.

Wu et al. (2022) pointed out that GBAs may capture unintended constructs instead of, or in addition to, those they intend to assess. More precisely, in their study, Wu et al. found that a GBA intending to measure consciousness captured cognitive ability instead. However, if our GBA is in fact a good measure of cognitive ability, then it should not correlate, or correlate only slightly, with tests intended to measure other constructs such as personality traits. Thus, drawing on the finding and the warning given by Wu et al., we wanted to ensure that performance in our GBA would not inadvertently be driven by unintended constructs such as test takers' personality, as personality traits, such as the Big Five, typically only show low correlations with scores in traditional cognitive ability tests (Beauducel et al., 2007; Liepmann et al., 2012; Schilling et al., 2020). Therefore, we also explored correlations between GBA performance and the Big Five and attempted to answer the following research question:

Research Question 1: *How strongly does GBA performance correlate with the Big Five personality factors?*

Our expectation was that the correlations of GBA performance with the Big Five would at least be significantly lower than the correlation of GBA performance with the traditional cognitive ability test, as our GBA indented to measure cognitive ability. Furthermore, we assumed that the correlations between GBA performance and the Big Five would all be relatively small as research indicated that cognitive ability only slightly correlates with personality (Beauducel et al., 2007; Liepmann et al., 2012; Schilling et al., 2020). Thus, this would represent additional evidence for the construct-related validity of the GBA.

3.2 | Applicant reactions toward GBAs

In recent years, owing to the war for talent, more and more attention is being paid not only to the validity of selection instruments, but also to applicant reactions. This is because organizations benefit relatively little if their selection process is valid but is not well accepted by applicants. In fact, applicants may reject a potential job offer if they perceive a selection tool as unfair or if they make other negative experiences during the application process (Hausknecht et al., 2004). Additionally, applicant perceptions might also impact applicants' test motivation, which is an antecedent for test performance (Hausknecht et al., 2004; Sanchez et al., 2000; Truxillo et al., 2009). Accordingly, it is important to consider applicant reactions when developing a novel selection instrument. Therefore, another goal of the present study was to investigate different applicant reaction variables toward our GBA, as empirical evidence in this area is also relatively limited to date (Woods et al., 2020). Several reviews and chapters on gamification and GBAs postulated that one of the main advantages of GRAs is their positive influence on applicants' reactions (see Bhatia & Ryan, 2018; Fetzer et al., 2017, and Landers et al., 2018, for detailed overviews). Accordingly, as described above, several prior studies in this area indicated that gamified assessments and GDAs

may lead to more positive applicant reactions compared with their traditional counterparts (e.g., Georgiou & Nikolaou, 2020; Gkorezis et al., 2020; Hommel et al., 2022; but also see Ohlms et al., 2023, or Georgiou, 2021). However, further research is needed, especially on GBAs assessing cognitive ability, to draw sound conclusions on whether and under what circumstances cognitive GBAs are better accepted by test takers than traditional ability tests. Therefore, it is also important to test whether certain individual characteristics (e.g., age, video game experience) may be positively or negatively related to applicant reaction variables. This would allow us to draw conclusions for which groups of potential applicants the use of GBAs may have a particularly positive effect (Ramos-Villagrasa et al., 2022). In addition, if applicant reactions to the present GBA differ from reactions to previous GBAs, this could be a starting point for future research on the effects of specific design elements (e.g., game genre, avatars, levels). This in turn might contribute to practitioners getting a better idea of which design elements they might want to use when developing a GBA and which should rather be left untouched (Ramos-Villagrasa et al., 2022).

Gilliland's (1993) fairness model is the most influential model in the field of applicant reactions (McCarthy et al., 2017). Among other things, this model postulates 10 rules of procedural fairness that are supposed to influence the perceived fairness of the selection process among applicants. Empirically, procedural fairness is substantially positively correlated with applicants' behavioral intentions (e.g., the intention to accept a job offer) and attitudes toward the organization (Hausknecht et al., 2004).

Based on the above-mentioned empirical findings and Gilliland's (1993) fairness model, there are several aspects of GBAs that may offer a promising opportunity to positively influence applicants' reactions (Fetzer et al., 2017; Woods et al., 2020): The interactive nature of video games grants users a degree of autonomy in completing them. For example, there might be a possibility to move freely in the environment within a GBA using the corresponding control functions of the respective game or to freely determine the order in which tasks are completed. In contrast, this is hardly feasible using traditional general mental ability (GMA) or personality tests. Related to this, it has also been argued that the increased degree of flexibility in completing a GBA compared with a traditional test should result in a higher perceived opportunity to perform (Landers et al., 2022), which is an important aspect of procedural fairness according to Gilliland's (1993) fairness model. Furthermore, meta-analytic results indicate that the perceived opportunity to perform is strongly related to procedural fairness as well as to other fairness aspects such as face validity and perceived predictive validity, and with various outcomes such as organizational attractiveness, recommendation intentions, and job offer acceptance intentions (Hausknecht et al., 2004).

In the context of GBAs, Ellison et al. (2020) found that different aspects of procedural fairness rules were positively associated with the perceived overall fairness of the selection process which, in turn, correlated positively with the willingness to recommend the organization to others. Additionally, Landers et al. (2022) found that their GBA was perceived more positively in terms of procedural fairness, distributive fairness, job relatedness, and test propriety compared with a paper-pencil test. Furthermore, al-Qallawi and Raghavan (2022) examined user reviews and comments from 10 GBA applications on two mobile distribution platforms. Generally, they found many positive evaluations from applicants, who particularly perceived GBAs as innovative. Nevertheless, there was also a substantial number of negative comments (often related to technical issues) concerning aspects that might impair applicant reactions.

Another potential advantage of GBAs is that the actual assessment of applicants' knowledge, skills, abilities, and other characteristics might be less transparent by being directly embedded in the game environment (Shute & Ventura, 2013). Thus, assessing cognitive ability through a GBA rather than a paper-pencil test might offer the opportunity to reduce test anxiety, as the salience of the test situation might be lower for applicants (Bhatia & Ryan, 2018). In the context of personnel selection, reductions in test anxiety might be relevant as meta-analytic evidence shows that anxiety is associated with lower levels of test motivation and test attitudes (Hausknecht et al., 2004). These, in turn, directly affect procedural fairness perceptions, job offer acceptance intentions, and recommendation intentions (Hausknecht et al., 2004). In line with this potential advantage of GBAs, several studies from the educational context found that GBAs reduced test anxiety compared with traditional exams (e.g., Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011). Furthermore, an initial study examining the use of a GBA to measure cognitive ability in the personnel selection and assessment context also indicated lower levels of test anxiety compared with a traditional intelligence test (Landers et al., 2022).

Taken together, the theoretical arguments based on Gilliland's justice model (1993) and the empirical evidence reviewed above lead to the following hypothesis:

Hypothesis 2: Applicants' reactions are more positive for the GBA than for a traditional cognitive ability test.

4 | METHOD

4.1 | Sample

We conducted an a priori power analysis to determine the required sample size to test our hypotheses with a power of 0.80. The analysis for Hypothesis 1, which predicted a large correlation (i.e., $r = .50$ according to common standards; cf. Cohen, 1988), revealed a required sample size of $N = 26$ for a two-tailed bivariate correlation. Additionally, for the power analysis related to Hypothesis 2, we assumed an effect size of $d = .25$ on the basis of Landers et al. (2022). This analysis revealed a sample size of $N = 128$ for two-tailed paired sample t-tests as well as for the within-group main effect in 2×2 mixed analysis of variance (ANOVAs).

We tested a sample of working individuals that initially consisted of 202 Bachelor graduates of the Police Academy of

Lower Saxony in Germany who voluntarily participated in this study. Due to technical issues with the GBA server, scores from only 158 participants were available for the GBA. This *N* is larger than the required sample size found in our power analyses. In addition, no scores for the paper-pencil test were available from one participant, as they did not hand in their answer sheet. Furthermore, data on applicant reactions to the GBA was available from 175 participants and for applicant reactions to the paper-pencil test from 188 participants. Data on personality, video game experience, and demographic variables were available from 193 participants. This was because some participants did not answer the questionnaires, or their data had to be removed from the data set as careless responding to the questionnaires had to be assumed due to their short completion time (i.e., duration <180 s for the applicant reactions questionnaires and <270 s for the questionnaire measuring personality, video game experience, and demographic variables). Since the results of the GBA and the paper-pencil test as well as participants' personality scores were relevant for the hypothesis and research question on validity, only data from participants with no missing values concerning these variables were used for the corresponding analyses. Moreover, analyses testing Hypothesis 2 were based solely on data from participants with complete data from all three questionnaires ($n = 156$). The total number of participants, whose data could be used for the analyses, consisted of $N = 183$.

Within the sample, age ranged from 20 to 34 with a mean of 22.83 years ($SD = 2.43$), and 43.7% of the participants were female and 56.3% were male. All participants had a Bachelor's (97.8%) or a master's degree (2.2%). The majority did not play video games on a weekly basis (54.6%).

4.2 | Procedure

Before the main data collection, we conducted a small pilot study with nine participants to examine clarity of the instructions, difficulty of the test, and applicant reactions. The results of the pilot study showed a high variability in the GBA and paper-pencil test scores, implying an appropriate level of difficulty.

In the main study, participants' informed consent was obtained first. Then they were instructed to imagine having applied for their current position once again and that they were now invited to an approximately 90-min recruitment event at the organization as part of the selection process.

Participants completed the GBA and, directly afterward, answered an online questionnaire that assessed their reactions regarding the GBA. Similarly, after they had taken the paper-pencil GMA test, participants completed a questionnaire assessing their reactions to it. Finally, information on participants' personality and demographic data was collected. To control for possible order effects, participants randomly completed the GBA or the paper-pencil test first.

4.3 | Measures

4.3.1 | Game-based assessment

As mentioned above, the GBA was implemented in the development environment of Minecraft (an excerpt of the GBA can be found at https://osf.io/qvw4z/?view_only=d0a1d22208184dad81223edef4403212). At the beginning of the GBA, participants received an explanation of the control buttons (they had to use the arrow keys, mouse, and spacebar to navigate in the GBA) and were informed about the cover story in which the assessment was embedded. Specifically, participants were instructed that they had to release a country from a curse by finding a crystal hidden on top of a multistory tower. Inside this tower, 36 items had to be completed to finish the game. These items aimed to assess cognitive ability by targeting the second-stratum ability fluid intelligence. They were chosen to reflect sequential, quantitative, and inductive reasoning. Specifically, the game contained 12 items each for verbal, numerical, and figural intelligence, which together formed the total cognitive ability score. Kuder–Richardson (KR-)20 across all 36 items of the GBA was $\alpha = .58$.

Once the participants had completed all the items, they reached the top of the tower. There they found the crystal to free the country from the curse and to successfully complete the GBA. Regardless of how many items were solved correctly, participants reached the top of the tower when they had completed all items. The overall GBA score was determined by the number of correctly solved items. One point was awarded for each correctly solved item, meaning GBA scores could range between 0 and 36.

Concerning the specific items, several boards were placed on the walls of the first and second floors of the tower with numerical items (for an example item see Figure 1a). On each of these boards, number sequences were displayed, and participants had to find the subsequent number in the series of numbers. There was a time limit of 6 min for solving the number sequences on the first floor of the tower (six items) and 4 min and 45 s for the corresponding items on the second floor (six items). More time was allowed on the first floor to give participants the opportunity to familiarize themselves with the controls. KR-20 for the numerical items was $\alpha = .63$. Participants gave the answer to an item by clicking on a corresponding button, which was located on the wall.

On the next two floors of the tower, verbal ability was targeted (for an example item see Figure 1b). Again, several boards were attached to the walls, each presenting three words and five answer options. There was always a specific relationship between the first two words, and participants had to select the option that had a similar association with the third word as the first two words had with each other (e.g., “taste and tongue are related to each other like smell is related to a. stink, b. nose, c. fragrance, d. breath"). At each of the two floors (six items each), 3 min and 30 s were given for completion. KR-20 for this subscale was 0.20.

Finally, to measure figural ability, different matrices items were designed, located on two floors of the tower (for an example item see

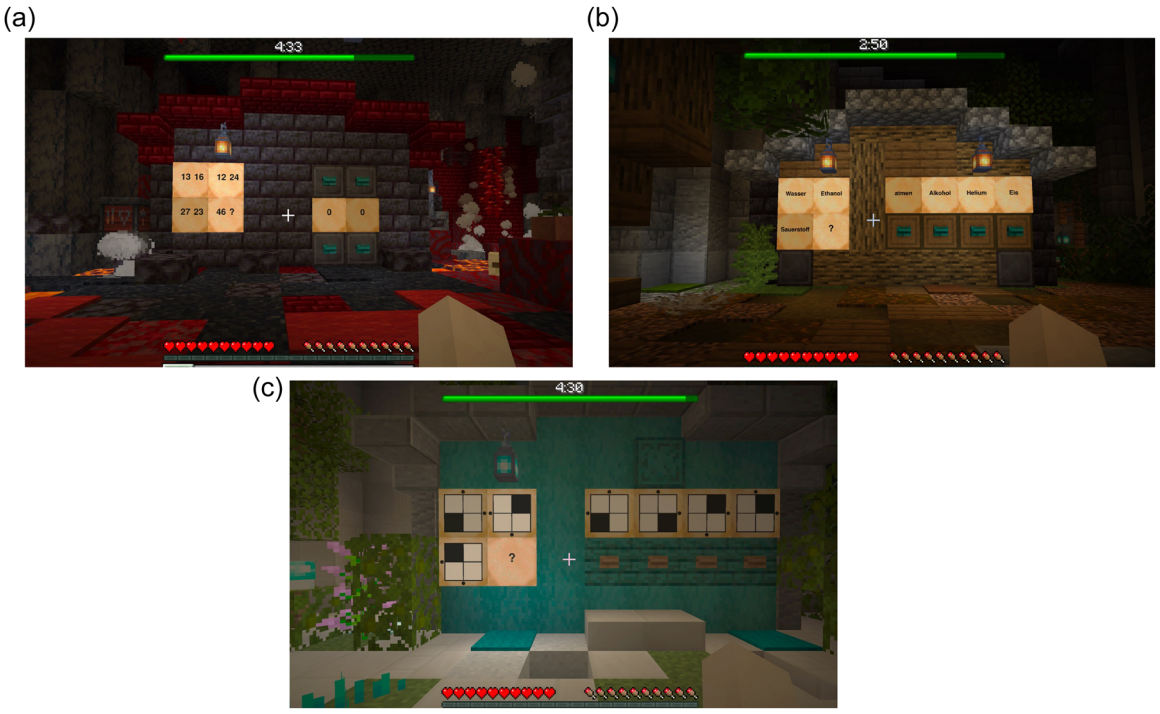

FIGURE 1 Example items for the three different item groups within the GBA. (a) Example item for measuring numerical intelligence. (b) Example item for measuring verbal intelligence. (c) Example item for measuring figural intelligence. GBA, game-based assessment.

Figure 1c). Within each item, three figures were presented following a certain pattern as well as five answer options. For the figural items, 4 min and 45 s were allowed per floor (six items each). KR-20 was 0.40.

In addition to the GMA items, the GBA also included two mini-games that did not affect the overall GMA score. The goal of integrating these two mini-games was to increase the fun and entertainment value of the GBA for the test takers. In the first mini-game, chickens, each of which appeared only briefly, had to be hit with snowballs by clicking on them with the mouse. In the second mini-game, test takers had to try to cross a room by jumping from one block to another (pressing the space bar to jump) without falling off of a block and onto the floor.

Only the number of correctly solved items counted toward the total score in the GBA, but no further behavior of participants within the GBA did (e.g., performance in the mini-games, time spent in a specific level, walking around the tower). The participants were free to choose the order in which they solved the items on one specific level.

4.3.2 | Paper-pencil cognitive ability test

The Intelligence-Structure-Test Screening (IST Screening; Liepmann et al., 2012) was used as the paper-pencil GMA test. This test is commonly used in Germany and consists of three groups, containing 20 items each, assessing verbal (i.e., analogies), numerical (i.e., number sequences), and figural ability (i.e., matrices). This test measures the same three Stratum I abilities of the CHC model as our GBA. Thus, the three task domains targeted in the IST Screening were chosen to best reflect the three key item categories intended to be measured by the GBA. The three subscales of the IST Screening can be combined to form an overall deductive reasoning score (KR-20 = 0.69).

4.3.3 | Personality

The Big Five were measured using the German version of the NEO-FFI (Borkenau & Ostendorf, 2008), which contains 10 items for each trait. A 5-point rating scale ranging from 1 = *strongly disagree* to 5 = *strongly agree* was used for these and the subsequent variables. Cronbach's α for the five scales ranged from .77 (agreeableness) to .85 (conscientiousness).

4.3.4 | Applicant reactions

Applicant reactions were obtained for the GBA and the paper-pencil test. Our goal was to collect a broad range of different applicant reaction variables to comprehensively capture reactions to the GBA and the paper-pencil test. Thus, based on a comprehensive literature review, we selected seven applicant reaction variables. The only modifications made to the scales were to ask either about the "test" or the "computer game". For the current study, scales for which no German version was available were translated into German and

checked with back-translation (see Table 2 for reliabilities of the different scales).

We used adapted items from Oostrom et al. (2013) that originally stem from Smither et al. (1993) to measure perceived face validity (three items, e.g., "The actual content of the computer game/test is related to the tasks of my job") and perceived predictive validity (three items, e.g., "The employer can tell a lot about the applicant's ability to do the job based on the results of the test"). Furthermore, two subscales from the German translation by Manzey and Gurk (2005) of the Selection Procedural Justice Scale (SPJS; Bauer et al. (2001), were used to measure perceived opportunity to perform (four items, e.g., "I could really show my skills and abilities through this computer game/test"), and overall procedural fairness (three items, e.g., "Overall, the method used was fair"). In addition to perceptions of the GBA and the test, organizational attractiveness was measured using the 5-item general attractiveness subscale (e.g., "For me, this company would be a good place to work") from the organizational attractiveness scale from Highhouse et al. (2003) and behavioral intentions were measured with four items from the pursuit intentions subscale (e.g., "I would accept a job offer from this company") from the same instrument. For both, the German translation from Basch et al. (2022) were used. Finally, test anxiety was measured using an adapted 6-item scale (e.g., "I usually get very anxious about taking tests") from Arvey et al. (1990) that was also previously used in other studies (e.g., Landers et al., 2022).

4.3.5 | Other variables

Experience with video games was measured using the 6-item scale from Bourgonjon et al. (2010; e.g., "I often play video games"). In addition, participants were asked about their video game habits by indicating how many hours per week they play video games in general and Minecraft specifically. Furthermore, demographic variables such as age, sex, and highest educational degree were assessed.

5 | RESULTS

5.1 | Preliminary analysis

Due to the considerable number of participants with missing data from the GBA, the paper-pencil test, and/or the other questionnaires, analyses were conducted to investigate whether these participants differed from those without missing data. To do so, we tested whether participants, who had no GBA scores due to technical problems, differed from those with scores on various variables (i.e., conscientiousness, age, sex). In addition, we examined whether participants, who were excluded because of a too-short completion time or who denied answering the questionnaires or the cognitive ability test, differed from the other participants. For both groups (i.e., missingness due to technical problems or too short response time and refusal to answer), no differences were found for conscientiousness, age, GBA scores, and IST Screening scores (all $ts \leq 1.51$, all $ps \geq .13$) or sex (all $\chi^2 s \leq 2.19$, all $ps \geq .14$).

Means, standard deviations, and correlations for the most relevant study variables are shown in Table 1. Unexpectedly, sex correlated with the overall GBA ($r = -0.31$, $p < .001$) and IST score ($r = -0.30$, $p < .001$) with males scoring higher than females. In addition, video game experience correlated positively with the overall GBA ($r = .33$, $p < .001$) and IST test scores ($r = .29$, $p < .001$).

5.2 | Tests of hypotheses

5.2.1 | Construct-related validity of the GBA

First, we examined correlations between the GBA and the IST. Evidence of convergence between the GBA and the IST would provide empirical support for the construct-related validity of the GBA (Hypothesis 1). To do so, Pearson product-moment correlations were calculated between overall performance on the GBA and the IST Screening, as well as between the three specific fluid abilities (i.e., verbal, numerical, and figural reasoning; see Table 1). Results indicated a strong positive correlation between the overall GBA and paper-pencil test scores of $r = .51$, $p < .001$. Similarly, the corresponding Stratum I abilities correlated positively with each other. Thus, for numerical and figural ability, medium-sized positive correlations of $r = .35$ and .36, both $ps < .001$, were found between the GBA and the IST Screening, and the verbal subscales of the two instruments correlated with $r = .18$, $p = .03$. Furthermore, the numerical and figural subscales of the GBA showed the strongest correlations with the corresponding IST Screening subscales. Given the strong positive correlation between the overall GBA and paper-pencil test score, Hypothesis 1 was supported for the overall GBA score but not for the three subtests, as there were only show small to medium-sized correlations between the IST and GBA subscales.

Additionally, to check whether our GBA was unintendedly measuring personality (see Wu et al., 2022), we tested whether GBA performance correlated with the Big Five (Research Question 1). We found that the Big Five neither correlated with the overall GBA score, nor with the overall IST Screening score (all $rs < .16$, see Table 1). Next, we compared the correlation between GBA and paper-pencil test performance statistically with the correlations of GBA performance with the individual Big Five scores. Therefore, a series of one-tailed Steiger's (1980) tests for dependent groups with a common third variable were conducted using the online tool from Hemmerich (2017). The results consistently showed that the correlation between the GBA and the IST Screening scores was significantly higher than between the GBA score and agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience, all $zs > 3.86$, all $ps < .001$. Thus, addressing Research Question 1, our

TABLE 1 Descriptive information and correlations for study variables.

Variables	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Age	22.83[a]	2.43[a]	–															
2. Sex	0.56[a]	0.50[a]	−0.06	–														
3. VG experience	2.12[a]	1.24[a]	−0.03	−0.77**	(0.95)													
4. GBA overall	21.96	3.67	−0.12	−0.31**	0.33**	(0.58)												
5. GBA verbal	9.48	1.46	−0.02	−0.30**	0.28**	0.50**	(0.20)											
6. GBA numeric	6.36	2.17	−0.09	−0.21**	0.24**	0.77**	0.08	(0.63)										
7. GBA figural	6.12	1.83	−0.13	−0.13	0.16*	0.71**	0.10	0.29**	(0.40)									
8. IST overall	47.03	4.83	−0.13	−0.30**	0.29***	0.51**	0.24**	0.37**	0.38**	(0.69)								
9. IST verbal	15.86	1.70	0.09	−0.22**	0.20*	0.24**	0.18*	0.20*	0.10	0.57**	(0.21)							
10. IST numeric	16.23	2.57	−0.20*	−0.29**	0.26**	0.41**	0.14	0.35**	0.30**	0.76**	0.18*	(0.70)						
11. IST figural	14.93	2.48	−0.11	−0.13	0.15	0.40**	0.21**	0.23**	0.36**	0.77**	0.23**	0.33**	(0.53)					
12. Agreeableness	3.90	0.45	−0.03	0.27**	−0.23**	−0.08	−0.21*	−0.01	0.02	−0.04	0.00	−0.12	0.04	(0.77)				
13. Conscientiousness	3.92	0.53	0.04	0.12	−0.25**	−0.03	−0.11	0.02	0.01	−0.08	−0.05	−0.10	−0.02	0.19*	(0.85)			
14. Extraversion	3.56	0.48	0.13	0.14	−0.14	−0.08	−0.21**	0.00	0.01	−0.15	−0.12	−0.15	−0.06	0.34***	0.22**	(0.78)		
15. Neuroticism	2.33	0.55	−0.19*	0.20*	−0.07	−0.06	0.00	−0.10	−0.01	−0.05	−0.04	−0.03	−0.03	−0.25**	−0.42**	−0.34**	(0.81)	
16. Openness	3.17	0.60	0.0	0.60	−0.03	0.10	0.08	0.06	0.06	−0.01	0.11	−0.10	−0.01	0.22**	−0.01	0.08	0.04	(0.78)

Note: $n = 151$. Scale reliabilities (internal consistencies) are presented on the diagonal, between parentheses. Sex is coded 0 = *male*, 1 = *female*.

Abbreviations: GBA, game-based assessment; IST, paper-pencil cognitive ability test; VG experience, video game experience.

[a]$N = 183$.

*$p < .05$; **$p < .01$; ***$p < .001$.

results do not suggest that the GBA inadvertently measures personality.

5.2.2 | Applicant reactions toward the GBA

To test Hypothesis 2, which predicted that applicants' perceptions would be more positive for the GBA than for the traditional cognitive ability test, various paired-samples *t*-tests were conducted to compare perceptions of the GBA and the IST Screening. Results indicated an opposite pattern than that predicted by Hypothesis 2. Thus, all applicant reaction variables were rated significantly more favorably for the paper-pencil test (see Table 2). According to common standards (Cohen, 1988), these effects were small to moderate for face validity, predictive validity, opportunity to perform, organizational attractiveness, behavioral intentions toward the organization, and test anxiety, but large for general procedural fairness. Altogether, Hypothesis 2 was clearly rejected.

5.3 | Further exploratory analyses

Additionally, we explored whether sex or video game experience were related to reactions toward the GBA and the paper-pencil test. To do so, a series of 2 × 2 mixed ANOVAs were conducted. The selection instrument (GBA vs. paper-pencil test) was entered as a within-subjects variable, and either sex or video game experience was used as a between-subjects variable. Before the analyses, video game experience (rated on a 5-point scale from 1 to 5) was dichotomized by using a median split. With a sample size of $N = 156$, we had a power of 0.99 to detect a medium-sized within-between interaction in these ANOVAs and a power of 0.70 to detect a small within-between interaction.

In the two ANOVAs, main effects of the selection instrument were still found for all applicant reaction variables, with higher ratings for the paper-pencil test than for the GBA (see Tables 3 and 4). For sex, there was a significant main effect for face validity, general procedural fairness, opportunity to perform, behavioral intentions,

TABLE 2 Reactions toward the GBA compared with the paper-pencil cognitive ability test.

	GBA			Paper-pencil test						95% CI for *d*	
Variable	*α*	*M*	*SD*	*α*	*M*	*SD*	*t*(155)	*p*	*d*	*LL*	*UL*
Face validity	.73	1.89	0.78	.80	2.18	0.83	−4.16	<0.001	−0.33	−0.49	−0.17
Predictive validity	.77	1.85	0.70	.76	2.13	0.73	−4.86	<0.001	−0.39	−0.55	−0.23
General procedural fairness	.77	2.53	0.91	.79	3.42	0.84	−11.46	<0.001	−0.92	−1.10	−0.73
Opportunity to perform	.78	1.99	0.84	.91	2.66	0.96	−8.41	<0.001	−0.67	−0.85	−0.50
Organizational attractiveness	.91	3.20	1.09	.87	3.82	0.80	−6.95	<0.001	−0.56	−0.72	−0.39
Behavioral intentions	.91	3.38	1.08	.87	3.99	0.80	−6.95	<0.001	−0.56	−0.72	−0.39
Test anxiety	.73	2.57	0.71	.76	2.21	0.66	5.70	<0.001	0.46	0.29	0.62

Note: $N = 156$. *t*-Tests and Cohen's *d* are calculated for paired comparisons. Positive *t*s and *d*s indicate larger scores for the GBA.
Abbreviations: GBA, game-based assessment; LL, lower limit; UL, upper limit.

TABLE 3 Results of two-way mixed ANOVAs for applicant reactions as dependent variables and sex as between-subject variable.

	GBA				Paper-pencil test				ANOVA					
	Males		Females		Males		Females		Sex		Selection instrument		Interaction	
Dependent variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*(1, 154)	η^2_p	*F*(1, 154)	η^2_p	*F*(1, 154)	η^2_p
Face validity	2.02	0.85	1.80	0.72	2.35	0.83	2.06	0.81	5.46*	0.03	17.49**	0.10	0.26	<0.01
Predictive validity	1.98	0.73	1.76	0.66	2.17	0.71	2.11	0.76	1.99	0.01	20.97**	0.12	1.88	0.01
Gen. proced. fairness	2.78	0.86	2.36	0.90	3.54	0.72	3.33	0.91	7.48**	0.05	123.21**	0.44	1.90	0.01
Opportunity to perform	2.31	0.85	1.77	0.76	2.79	0.85	2.57	1.03	10.28**	0.06	64.57**	0.30	4.10*	0.03
Organ. attractiveness	3.45	1.07	3.03	1.01	3.81	0.71	3.83	0.86	2.51	0.02	42.98**	0.22	5.99*	0.04
Behavioral intentions	3.72	0.92	3.14	1.12	4.00	0.73	3.99	0.85	5.80*	0.04	42.44**	0.22	11.01*	0.07
Test anxiety	2.28	0.62	2.78	0.69	2.05	0.59	2.32	0.69	20.40**	0.12	28.75**	0.16	3.19	0.02

Note: $n_{Male} = 65$; $n_{Female} = 91$. No variance homogeneity for perceptions of general fairness of the paper-pencil test and organizational attractiveness of the GBA. A Box-Cox transformation (Box & Cox, 1964) was used to stabilize the variances, which did not change the ANOVA results in terms of significance level, so that results could be interpreted despite missing variance homogeneity.
Abbreviations: GBA, game-based assessment; Gen. proced. fairness, general procedural fairness.
*$p < .05$; **$p < .01$.

TABLE 4 Results of two-way mixed ANOVA for applicant reactions as dependent variables and video game experience as between-subject variable.

	GBA				Paper-pencil test				ANOVA					
	Low VG exp.		High VG exp		Low VG exp.		High VG exp.		VG experience		Selection instrument		Interaction	
Dependent variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*(1, 154)	η^2_p	*F*(1, 154)	η^2_p	*F*(1, 154)	η^2_p
Face validity	1.84	0.73	1.94	0.83	2.01	0.74	2.33	0.88	3.67	0.02	16.98**	0.10	2.73	0.02
Predictive validity	1.76	0.69	1.94	0.70	2.01	0.71	2.25	0.74	4.56*	0.03	23.32**	0.13	0.19	<0.01
Gen. proced. fairness	2.35	0.87	2.70	0.91	3.30	0.88	3.53	0.79	6.35*	0.01	131.42**	0.46	0.60	<0.01
Opportunity to perform	1.73	0.76	2.23	0.84	2.51	1.05	2.80	0.85	11.19**	0.07	71.84**	0.32	0.18	0.01
Organ. attractiveness	2.92	1.12	3.46	1.01	3.83	0.87	3.81	0.73	4.37*	0.03	53.00**	0.26	10.48**	0.06
Behavioral intentions	3.08	1.17	3.66	0.91	3.97	0.87	4.01	0.75	6.45*	0.04	52.72**	0.26	9.97**	0.06
Test anxiety	2.82	0.75	2.34	0.58	2.37	0.76	2.06	0.51	22.59**	0.13	33.23**	0.18	1.88	0.01

Note: $n_{\text{VG exp. <1.6}} = 75$; $n_{\text{VG exp. ≥1.6}} = 81$. No variance homogeneity for perceptions of general fairness and text anxiety of the paper-pencil test as well as organizational attractiveness and test anxiety of the GBA. A Box-Cox transformation (Box & Cox, 1964) was used to stabilize the variances, which did not change the ANOVA results in terms of significance level, so that results could be interpreted despite missing variance homogeneity. No equality of covariances for the 2 × 2 mixed ANOVA with test anxiety toward the GBA and paper-pencil test as within-subjects factors.

Abbreviations: GBA, game-based assessment; Gen. proced. fairness, general procedural fairness; High VG exp., video game experience ≥1.6; Low VG exp., video game experience <1.6; VG experience, video game experience.

*$p < .05$; **$p < .01$.

and test anxiety, with males showing more positive reactions and reporting less test anxiety. In addition, participants with higher video game experience rated predictive validity, general procedural fairness, opportunity to perform, organizational attractiveness, and behavioral intentions more positively across both selection instruments, and reported lower test anxiety. Finally, three significant interaction effects were found between sex and the selection instrument. These concerned opportunity to perform, $F(1, 154) = 4.10$, $p = .046$, $\eta^2_p = .03$, perceived organizational attractiveness, $F(1, 154) = 5.99$, $p = .02$, $\eta^2_p = .04$, and behavioral intentions toward the organization, $F(1, 154) = 11.01$, $p = .001$, $\eta^2_p = .07$. For these variables, differences between men and women were larger for the GBA than for the paper-pencil test. Regarding video game experience, two interaction effects were found for organizational attractiveness, $F(1, 154) = 10.48$, $p = .001$, $\eta^2_p = .06$, and behavioral intentions, $F(1, 154) = 9.97$, $p = .002$, $\eta^2_p = .06$. For these variables, differences between participants with low versus high video game experience were larger for the GBA than for the paper-pencil test.

6 | DISCUSSION

The present study makes several important contributions to the literature. First, it responds to previous calls for further studies on GBAs, particularly regarding their validity, as well as applicant reactions (e.g., Ramos-Villagrasa et al., 2022; Woods et al., 2020). Specifically, to the best of our knowledge, it is the first study that has examined the psychometric properties (i.e., validity) of a GBA from the adventure genre intending to measure cognitive ability for use in personnel selection. Second, it examined whether applicant reactions toward this GBA differed from reactions to a traditional paper-pencil test. And third, we found evidence that the different applicant reaction variables were related to participants' sex as well as to their previous video game experience.

Concerning the first contribution, we examined whether a GBA can be used for personnel selection purposes to measure applicants' cognitive abilities if its development is theory-guided. On the one hand, the individual items within the GBA were developed based on the contemporary intelligence model, the CHC model (McGrew, 2009), so that performance in the GBA should converge with performance in a traditional fluid intelligence test to the greatest possible extent. On the other hand, when selecting and programming the individual game elements, much attention was paid to the existing literature on GBAs and gamification (e.g., Bedwell et al., 2012; Landers et al., 2018). Importantly, we found that overall performance on the present GBA correlated strongly with overall performance on a paper-pencil ability test ($r = .51$) that intended to measure the same cognitive abilities, but not with the Big Five personality factors. This pattern of results suggest that GBAs can be designed in such a way that they allow for a relatively construct valid measurement of cognitive ability, without inadvertently measuring other unintended constructs (i.e., personality). However, for the subtests of the GBA and the traditional cognitive ability test, we found only moderate correlations for numerical and figural intelligence and a small correlation for verbal intelligence. This argues for using the overall GBA score rather than the subtests when interpreting participants' performance. These results are in line with previous research (e.g., Barends et al., 2021; Georgiou et al., 2019; Landers et al., 2022), which demonstrated that theory-driven development of GRAs can produce valid personnel selection instruments.

Furthermore, our study provides valuable insights for research and practice beyond previous research. For example, while Landers et al. (2022) chose a similar study design to ours (i.e., developing a GBA to measure cognitive abilities), their GBA was rather different from the GBA used in our study. Specifically, the GBA from Landers et al. (2022) intended to measure other specific abilities (quantitative knowledge, reading and writing, fluid reasoning, and processing speed) and used different mini-games that stem from the puzzle genre, whereas we targeted fluid intelligence and used an adventure game. Thus, the two GBAs differ considerably regarding the cognitive abilities measured as well as their design and game genre. Accordingly, our results provide further evidence that GBAs can be designed in such a way that they allow for convergence with traditional cognitive ability tests, and hence provide empirical support for the construct-related validity of GBAs designed to measure cognitive ability, and that this is possible across different cognitive abilities, game genres, and designs.

Concerning our second contribution regarding applicant reactions, an unexpected pattern of results emerged. In contrast to the often postulated assumption that GRAs are generally associated with more positive applicant reactions compared with traditional instruments (e.g., Bhatia & Ryan, 2018; Woods et al., 2020), this was not the case for the present GBA. More precisely, all reactions toward the traditional GMA test were more favorable than those toward the GBA with effect sizes ranging from small to large depending on the specific dependent variable. Previous studies that examined applicant reactions to GRAs often found more positive reactions to GRAs compared with their traditional counterparts (e.g., Georgiou & Nikolaou, 2020; Gkorezis et al., 2020; Hommel et al., 2022; Landers & Collmus, 2022; Landers et al., 2022). However, there are a few studies that found no differences concerning applicant reactions toward GRAs (e.g., Landers et al., 2020), or even more negative reactions to GRAs compared with non-gamified methods (e.g., Georgiou, 2021). Obviously, no two GBAs are alike, but the present results clearly suggest that GBAs are not per se perceived more positively. Instead, it seems that design elements (e.g., game genre, multimedia-style, cover story) can influence the appearance of gamified assessments, GDAs, and/or GBAs and that this may also influence how they are perceived by potential applicants (Fetzer et al., 2017).

Interestingly, our results also differed from those from Landers et al. (2022) who found more positive reactions to their cognitive-ability GBA. However, our GBA differed considerably from Landers et al.'s GBA regarding the design elements and game genre used. And even though we cannot say which differences in the game design elements and/or in the study design contributed to these differences, there are a few notable differences that might have played a role. First, our GBA belongs to the adventure genre, as, within the GBA, participants have to go on an adventure in a fictional world and are confronted with various problem-solving tasks. Second, our GBA is implemented in the Minecraft development environment, whereas Cognify consists of various puzzle-like tasks. In addition to differences concerning the two GBAs, the paper-pencil tests used in the two studies also differed considerably in their appearance. While the IST Screening (Liepmann et al., 2012), used in the current study, has a more modern design, the subtests used by Landers et al. (2022) tend to look more old-fashioned (see Ekstrom et al., 1976). Thus, the contrast between the innovative GBA and the traditional paper-pencil test was possibly more salient in the study by Landers et al. (2022) than it was in the present study. Unfortunately, however, it remains unclear to what extent the specific design elements of the GBAs and/or of the study design contributed to the different effects concerning the applicant reaction variables. Thus, more research that considers the potentially relevant aspects separately is necessary.

Concerning the third contribution, in line with earlier findings (Ellison et al., 2020; Melchers & Basch, 2022), we found evidence that participants' sex and video game experience were related to differences between the GBA and the paper-pencil test. Even though all participants (i.e., males vs. females and participants with high vs. low video game experience) showed more positive reactions toward the paper-pencil test compared with the GBA, we also found that men rated face validity, general procedural fairness, opportunity to perform, behavioral intentions, and test anxiety more positively than women across both instruments. Similarly, individuals with a larger deal of video game experience had more positive applicant reactions than participants with little video game experience across both instruments, except for face validity. Furthermore, significant interaction effects were found between the selection instrument and sex for opportunity to perform, perceived organizational attractiveness, and behavioral intentions toward the organization. Similarly, there was a significant interaction effect between the selection instrument and video game experience for perceived organizational attractiveness and behavioral intentions toward the organization. In all cases, differences between males versus females and between participants with low versus high video game experience were larger for the GBA than for the paper-pencil test; regarding reactions toward the paper-pencil test, there were hardly any differences between men and women or participants with high versus low video game experience.

Taken together, the interaction effects indicate the presence of subgroup differences in reactions to the GBA for organizational attractiveness and behavioral intentions, whereas these differences were absent for the traditional cognitive ability test. More specifically, women and individuals with low video game experience perceived an organization using the present GBA as less attractive and had more negative behavioral intentions toward such an organization than men and individuals with high video game experience. Hence, these findings suggest that the use of GBAs might differentially influence applicant reactions for different groups of applicants. Especially women and those less familiar with video games may have more negative attitudes toward this playful method than men and experienced video game players. However, it should also be noted that, for both selection instruments, reactions were relatively negative, with average scores across all reaction variables with means of $M = 2.92$ ($SD = 0.51$) for the paper-pencil test and $M = 2.49$ ($SD = 0.55$) for the GBA on a scale from 1 to 5. This may

reflect that GMA tests are generally less accepted than many other selection instruments (Anderson et al., 2010).

6.1 | Limitations and future research

Although this study provides important insights into the use of GBAs in the context of personnel selection, it is not without limitations. First, it should be emphasized that GBAs can take on very different appearances depending on their specific design (e.g., different game genres) and their purpose (e.g., assessment of personality vs. problem-solving skills; Bhatia & Ryan, 2018; Fetzer et al., 2017). Thus, the generalizability of the present results to other GBAs should be treated with caution. Nevertheless, this study showed that GBAs can be developed in such a way that they fulfill comparable psychometric standards, at least in terms of construct-related validity, as other more traditional instruments assessing the same constructs (also see Landers et al., 2022), and we assume that this can also be the case for other GBAs that incorporate relatively traditional intelligence items in a game. However, whether this is true or not has to be examined for each particular GBA. Furthermore, whether the relatively skeptical applicant reactions were driven by the content of the test items, the genre of the game, its visual design, or any other design feature is unclear so far. Hence, further research with other GBAs, measuring different constructs and using various design elements, is necessary to assess the generalizability of the obtained results and to gain more in-depth knowledge about the validity of GBAs and applicant reactions to them. According to Lievens and Sackett (2017) modular approach to personnel selection, rather than considering selection procedures as single entities, one should decompose them into their key elements and examine their effects. Following such a modular approach, to draw a solid conclusion about the effect of specific design elements (e.g., genre, avatars, levels) on the psychometric properties of and applicant reactions to a GBA, requires studying a specific design element in isolation to be able to attribute any changes to the gamification element. Thus, future research should examine the effects of specific design elements, such as using avatars or a storyline, on validity and reactions in isolation. In this respect, it might also be worthwhile to consider in isolation how the two mini-games used in our GBA might have influenced applicants' enjoyment and fun.

Another limitation of our GBA is that it used a static narrative (i.e., participants finished the game and saved the cursed country regardless of their actual performance). Accordingly, the overall GBA score was based on the number of correctly solved items, whereas participants' behavior within the game did not impact GBA performance or GBA scores. We chose this setting to keep the game mechanics as simple as possible while simultaneously increasing applicants' enjoyment of the assessment by embedding the actual items into a game. However, it would be interesting to investigate in a further development of the GBA whether participants' behavior in the game can also be validly integrated into the GBA score. Moreover, it might also be valuable to examine whether changing the game mechanics themselves would affect validity and applicant reactions, such as using a joystick instead of a keyboard to control the game.

Furthermore, the progress of the game did not depend on participants' performance and test takers also did not receive feedback on their performance. This was done to ensure that all participants were presented with the same items. However, whether or not test takers receive feedback could affect their motivation and thus their performance. Therefore, it would be valuable, for example, to explore whether giving or not giving feedback to participants on their performance affects their reactions and the validity of a specific GBA.

Furthermore, it should be noted that the overall GBA score as well as the verbal and numerical subscales correlated with sex, with males performing better than females. This is in line with previous concerns that males might be advantaged by GBAs (Bhatia & Ryan, 2018; Fetzer et al., 2017) and extends the limited research on subgroup differences in GBAs (Melchers & Basch, 2022). However, the same patterns of results were found for the traditional cognitive ability test. Thus, it could be that the GBA simply replicated the gender differences that are also apparent in the traditional test. Given that there are usually no meaningful sex differences for the IST Screening (Liepmann et al., 2012), it might be that the present sample was not representative. Future research should therefore re-examine sex differences in the GBA and the paper-pencil test, because potential subgroup differences might be problematic in the personnel selection context, as even small differences could lead to disadvantages for specific groups of applicants.

In addition, GBA performance was also associated with video game experience so that individuals with more video game experience scored better across all GBA scales than individuals with limited video game experience. This might be due to the fact that certain game mechanics such as manipulating objects (e.g., clicking answers or throwing), finding one's way around the game environment, or interpreting features of the game (e.g., what is the object in front of me and what can I do with it?) may be more challenging for players without game experience than for those with experience, and thus might represent an additional cognitive load for unexperienced test takers (Arthur et al., 2018). Thus, it is possible that the more complex the game mechanics are designed, the more difficult it is for non-gamers to understand them, which might put them at a disadvantage. Within our GBA, we attempted to design the game mechanics to be as easy to understand as possible, however, the cognitive load may have been higher for inexperienced gamers than for experienced gamers. Thus, future research should investigate how game mechanics can be designed in such a way that video game experience does not lead to an unfair advantage in the GBA.

Another limitation worth mentioning is that all participants were employed by the same government agency, meaning it cannot be ruled out that satisfaction with their current job and the organization, in general, may have biased the results, especially with regard to applicant reactions. Furthermore, given that all participants had at

least a Bachelor's degree, this may have led to some range restriction, for instance in terms of cognitive ability. This is because, in Germany, only those who have completed their Abitur (i.e., the final school leaving examination in Germany that qualifies for university admission; completed successfully by about half of all school leavers) can pursue a university degree. And given that academic performance is correlated with cognitive ability (Kuncel & Hezlett, 2007; Neisser et al., 1996), only people with a higher level of GMA tend to attend university, which is why there was potential restriction of variance in participants' cognitive ability. Accordingly, relationships between the two selection instruments might represent conservative estimates. Finally, we only conducted a single study, which is another reason why it would be interesting to replicate the present results in another, more diverse sample. This would allow for an evaluation of the replicability of the present results and for a more detailed exploration of potential effects of age, occupational field, and educational level on test performance and applicant reactions.

6.2 | Practical implications

In recent years, the number of commercial providers of digital selection methods, which include GRAs, as well as the number of organizations using GBAs for personnel selection, has increased rapidly (Armoneit et al., 2020; Woods et al., 2020). At the same time, there is still a lack of empirical research concerning the construct- and criterion-related validity of game-related methods (Woods et al., 2020; Ramos-Villagrasa et al., 2022). Addressing the scientist-practitioner gap, the results of the present study indicate that GBAs intending to measure cognitive abilities can indeed show a strong positive correlation with performance in a traditional paper-pencil ability test targeting the same abilities. However, the results of the present GBA are not simply generalizable to all other gamified or game-based instruments, since the different specific design elements and the particular construct(s) measured by a certain GBA can influence its psychometric properties (Bhatia & Ryan, 2018; Fetzer et al., 2017). Accordingly, the distinction between constructs and methods is essential when considering the use of GBAs (Arthur & Villado, 2008). Thus, the first step for organizations that want to implement GBAs in their selection process is to determine the constructs that they want to assess within their selection process. And it is only necessary in the second step to consider which of these constructs can or should potentially be measured using a GBA. However, given the sometimes limited support for the construct-related validity of GBAs concerning the targeted constructs (cf. Landers & Collmus, 2022) and the often-missing evidence concerning their criterion-related validity, any specific GBA should be validated before its implementation to ensure that it meets the standards of a psychometrically valid selection tool. Concerning the measurement of cognitive ability in GBAs, however, results from the present study as well as from previous research allow for optimism.

Furthermore, in times of the war for talent, applicant reactions are becoming increasingly important in attracting top talent to work for one's organization (Chambers et al., 1998). Contrary to suggestions that GBAs consistently lead to more positive applicant reactions compared with more traditional methods (Bhatia & Ryan, 2018; Fetzer et al., 2017), the present study found the opposite result. Thus, it cannot be taken for granted that GBAs are always more readily accepted by applicants than their traditional counterparts. One possibility to avoid potentially negative effects of a GBA on applicant reactions would be to obtain more precise information regarding its acceptance before its use in the actual selection process.

Altogether, before implementing a specific GBA in a selection process, it is necessary to evaluate whether the benefits of this novel method outweigh those of its traditional counterpart, taking into account multiple relevant outcomes. From an economic point of view, it should be considered, for example, whether the initially higher development costs of a GBA can save costs in the long term through automated implementation. In addition, outcomes such as validity, applicant reactions, and subgroup differences need to be carefully compared between GBAs and traditional selection methods.

ACKNOWLEDGMENTS

This work was partially supported by a doctoral scholarship of the Studienstiftung des Deutschen Volkes for Marie L. Ohlms. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Marie L. Ohlms http://orcid.org/0000-0003-4022-6904
Klaus G. Melchers http://orcid.org/0000-0003-4211-6450
Uwe P. Kanning http://orcid.org/0000-0002-9839-2940

REFERENCES

al-Qallawi, S., & Raghavan, M. (2022). A review of online reactions to game-based assessment mobile applications. *International Journal of Selection and Assessment*, *30*(1), 14–26. https://doi.org/10.1111/ijsa.12346

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*(3), 291–304. https://doi.org/10.1111/j.1468-2389.2010.00512.x

Ángeles Quiroga, M., Escorial, S., Román, F. J., Morillo, D., Jarabo, A., Privado, J., Hernández, M., Gallego, B., & Colom, R. (2015). Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can. *Intelligence*, *53*, 1–7. https://doi.org/10.1016/j.intell.2015.08.004

Armoneit, C., Schuler, H., & Hell, B. (2020). Nutzung, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020 [Use, validity, practicability, and acceptance of psychological personnel selection procedures in Germany 1985, 1993, 2007, 2020]. *Zeitschrift für Arbeits- und Organisationspsychologie*, *94*(2), 67–82. https://doi.org/10.1026/0932-4089/a000311

Arthur, Jr. W., Keiser, N. L., & Doverspike, D. (2018). An information-processing-based conceptual framework of the effects of unproctored internet-based testing devices on scores on employment-related

assessments and tests. *Human Performance*, *31*(1), 1–32. https://doi.org/10.1080/08959285.2017.1403441

Arthur, Jr. W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*(4), 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Baniqued, P. L., Lee, H., Voss, M. W., Basak, C., Cosman, J. D., DeSouza, S., Severson, J., Salthouse, T. A., & Kramer, A. F. (2013). Selling points: What cognitive abilities are tapped by casual video games. *Acta Psychologica*, *142*(1), 74–86. https://doi.org/10.1016/j.actpsy.2012.11.009

Barends, A. J., De Vries, R. E., & Van Vugt, M. (2019). Gamified personality assessment: Virtual behavior cues of Honesty- Humility. *Zeitschrift für Psychologie*, *227*(3), 207–217. https://doi.org/10.1027/2151-2604/a000379

Barends, A. J., de Vries, R. E., & van Vugt, M. (2021). Construct and predictive validity of an assessment game to measure honesty–humility. *Assessment*, *29*(4), 630–650. https://doi.org/10.1177/1073191120985612

Basch, J. M., Melchers, K. G., & Büttner, J. C. (2022). Pre-selection in the digital age: A comparison of perceptions of asynchronous video interviews with online tests and online application documents in a simulation context. *International Journal of Selection and Assessment*, *30*(4), 639–652. https://doi.org/10.1111/ijsa.12403

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the Selection Procedural Justice Scale (SPJS). *Personnel Psychology*, *54*(2), 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Beauducel, A., Liepmann, D., Felfe, J., & Nettelnstroth, W. (2007). The impact of different measurement models for fluid and crystallized intelligence on the correlation with personality traits. *European Journal of Psychological Assessment*, *23*(2), 71–78. https://doi.org/10.1027/1015-5759.23.2.71

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & Gaming*, *43*(6), 729–760. https://doi.org/10.1177/1046878112439444

Bhatia, S., & Ryan, A. M. (2018). Hiring for the win: Game-based assessment in employee selection. In J. H. Dulebohn, & D. L. Stone (Eds.), *The brave new world of eHRM 2.0* (pp. 81–110). Information Age Publishing.

Borkenau, P., & Ostendorf, F. (2008). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae [NEO Five-Factor Inventory according to Costa and McCrae]* (2nd ed.). Hogrefe.

Bourgonjon, J., Valcke, M., Soetaert, R., & Schellens, T. (2010). Students' perceptions about the use of video games in the classroom. *Computers & Education*, *54*(4), 1145–1156. https://doi.org/10.1016/j.compedu.2009.10.022

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Chambers, E., Foulon, M., Handfield-Jones, H., Hankin, S., & Michael, III E. (1998). The war for talent. *The McKinsey Quarterly*, *3*, 44–57.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology*, *9*(3), 621–640. https://doi.org/10.1017/iop.2016.6

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Collmus, A. B., & Landers, R. N. (2019). Game-framing to improve applicant perceptions of cognitive assessments. *Journal of Personnel Psychology*, *18*(3), 157–162. https://doi.org/10.1027/1866-5888/a000227

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification", In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). Association for Computing Machinery. https://doi.org/10.1145/2181037.2181040

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. Educational Testing Service.

Ellison, L. J., McClure Johnson, T., Tomczak, D., Siemsen, A., & Gonzalez, M. F. (2020). Game on! Exploring reactions to game-based selection assessments. *Journal of Managerial Psychology*, *35*(4), 241–254. https://doi.org/10.1108/JMP-09-2018-0414

Fetzer, M., McNamara, J., & Geimer, J. L. (2017). Gamification, serious games and personnel selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 293–309). Wiley. https://doi.org/10.1002/9781118972472

Georgiou, K. (2021). Can explanations improve applicant reactions towards gamified assessment methods? *International Journal of Selection and Assessment*, *29*(2), 253–268. https://doi.org/10.1111/ijsa.12329

Georgiou, K., Gouras, A., & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International Journal of Selection and Assessment*, *27*(2), 91–103. https://doi.org/10.1111/ijsa.12240

Georgiou, K., & Nikolaou, I. (2020). Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. *Computers in Human Behavior*, *109*, 106356. https://doi.org/10.1016/j.chb.2020.106356

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.5465/amr.1993.9402210155

Gkorezis, P., Georgiou, K., Nikolaou, I., & Kyriazati, A. (2020). Gamified or traditional situational judgement test? A moderated mediation model of recommendation intentions via organizational attractiveness. *European Journal of Work and Organizational Psychology*, *30*(2), 240–250. https://doi.org/10.1080/1359432X.2020.1746827

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hemmerich, W. (2017). *Korrelationen statistisch vergleichen* [Compare correlations statistically]. Statistik Guru. Retrieved October 6, 2021, from https://statistikguru.de/rechner/korrelationen-vergleichen.html

Highhouse, S., Lievens, F., & Sinar, E. F. (2003). Measuring attraction to organizations. *Educational and Psychological Measurement*, *63*(6), 986–1001. https://doi.org/10.1177/0013164403258403

Hommel, B. E., Ruppel, R., & Zacher, H. (2022). Assessment of cognitive flexibility in personnel selection: Validity and acceptance of a gamified version of the Wisconsin Card Sorting Test. *International Journal of Selection and Assessment*, *30*(1), 126–144. https://doi.org/10.1111/ijsa.12362

Kleinmann, M., & Strauß, B. (1998). Validity and application of computer-simulated scenarios in personnel assessment. *International Journal of Selection and Assessment*, *6*(2), 97–106. https://doi.org/10.1111/1468-2389.00078

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, *315*(5815), 1080–1081. https://doi.org/10.1126/science.1136618

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2022). Theory-driven game-based assessment of general cognitive

ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology*, *107*(10), 1655–1677. https://doi.org/10.1037/apl0000954

Landers, R. N., Auer, E. M., & Abraham, J. D. (2020). Gamifying a situational judgment test with immersion and control game elements: Effects on applicant reactions and construct validity. *Journal of Managerial Psychology*, *35*(4), 225–239. https://doi.org/10.1108/JMP-10-2018-0446

Landers, R. N., Auer, E. M., Collmus, A. B., & Armstrong, M. B. (2018). Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming*, *49*(3), 315–337. https://doi.org/10.1177/1046878118774385

Landers, R. N., & Collmus, A. B. (2022). Gamifying a personality measure by converting it into a story: Convergence, incremental prediction, faking, and reactions. *International Journal of Selection and Assessment*, *30*(1), 145–156. https://doi.org/10.1111/ijsa.12373

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment*, *30*(1), 1–13. https://doi.org/10.1111/ijsa.12376

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). I-S-T 2000 R: Intelligenz- Struktur-Test 2000 R *[I-S-T 2000 R: Intelligence Structure Test 2000 R.]* (2nd ed.). Hogrefe.

Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). IST Screening: Intelligenz-Struktur-Test Screening *[IST Screening: Intelligence Structure Test Screening]*. Hogrefe.

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology*, *102*(1), 43–66. https://doi.org/10.1037/apl0000160

Manzey, D., & Gurk, S. (2005, September 19–21). *Prozedurale Gerechtigkeit von Personalauswahlmaßnahmen: Untersuchungen zu einer deutschen Version der Selection Procedural Justice Scale (SPJS) von* Bauer et al. (2001) [Procedural justice of personnel selection methods: Investigations of a German version of the Selection Procedural Justice Scale (SPJS) by Bauer et al. (2001)] [Paper presentation]. 4th Annual Conference of the German Society for Work and Organizational Psychology, Bonn, Germany.

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. https://doi.org/10.1111/jcal.12170

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing "So what?," "What's new?," and "Where to next? *Journal of Management*, *43*(6), 1693–1725. https://doi.org/10.1177/0149206316681846

McCarthy, J. M., & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of Selection and Assessment*, *13*(4), 282–295. https://doi.org/10.1111/j.1468-2389.2005.00325.x

McChesney, J., Campbell, C., Wang, J., & Foster, L. (2022). What is in a name? Effects of game- framing on perceptions of hiring organizations. *International Journal of Selection and Assessment*, *30*(1), 182–192. https://doi.org/10.1111/ijsa.12370

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

McPherson, J., & Burns, N. R. (2007). Gs invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods*, *39*(4), 876–883. https://doi.org/10.3758/BF03192982

Melchers, K. G., & Basch, J. M. (2022). Fair play? Sex-, age-, and job-related correlates of performance in a computer-based simulation game. *International Journal of Selection and Assessment*, *30*(1), 48–61. https://doi.org/10.1111/ijsa.12337

Neisser, U., Boodoo, G., Bouchard, Jr. T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Nikolaou, I., Georgiou, K., & Kotsasarlidou, V. (2019). Exploring the relationship of a gamified assessment with performance. *The Spanish Journal of Psychology*, *22*, E6. https://doi.org/10.1017/sjp.2019.5

Ohlms, M. L., Voigtländer, E., Melchers, K. G., & Kanning, U. P. (2023). *Is gamification a suitable means to improve applicant reactions and convey information during an online test?* [Manuscript submitted for publication]. Institute of Psychology and Education, Ulm University.

Oostrom, J. K., van der Linden, D., Born, M. P., & van der Molen, H. T. (2013). New technology in personnel selection: How recruiter characteristics affect the adoption of new selection technology. *Computers in Human Behavior*, *29*(6), 2404–2415. https://doi.org/10.1016/j.chb.2013.05.025

Quiroga, M., & Colom, R. (2020). Intelligence and video games. In R. Sternberg (Ed.), *The Cambridge handbook of intelligence* (2nd ed., pp. 626–656). Cambridge University Press. https://doi.org/10.1017/9781108770422

Quiroga, M. A., Diaz, A., Román, F. J., Privado, J., & Colom, R. (2019). Intelligence and video games: Beyond "brain-games". *Intelligence*, *75*, 85–94. https://doi.org/10.1016/j.intell.2019.05.001

Quiroga, M. A., Herranz, M., Gómez-Abad, M., Kebir, M., Ruiz, J., & Colom, R. (2009). Video-games: Do they require general intelligence? *Computers & Education*, *53*(2), 414–418. https://doi.org/10.1016/j.compedu.2009.02.017

Quiroga, M. Á., Román, F. J., Catalán, A., Rodríguez, H., Ruiz, J., Herranz, M., Gómez- Abad, M., & Colom, R. (2011). Videogame performance (not always) requires intelligence. *International Journal of Online Pedagogy and Course Design*, *1*(3), 18–32. https://doi.org/10.4018/ijopcd.2011070102

Ramos-Villagrasa, P. J., Fernández-del-Río, E., & Castro, Á. (2022). Game-related assessments for personnel selection: A systematic review. *Frontiers in Psychology*, *13*:Article 952002. https://doi.org/10.3389/fpsyg.2022.952002

Ryan, A. M., & Derous, E. (2019). The unrealized potential of technology in selection assessment. *Revista de Psicología del Trabajo y de las Organizaciones*, *35*(2), 85–92. https://doi.org/10.5093/jwop2019a10

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040–2068. https://doi.org/10.1037/apl0000994

Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology*, *85*(5), 739–750. https://doi.org/10.1037/0021-9010.85.5.739

Schilling, M., Becker, N., Grabenhorst, M. M., & König, C. J. (2020). The relationship between cognitive ability and personality scores in selection situations: A meta-analysis. *International Journal of Selection and Assessment*, *29*(1), 1–18. https://doi.org/10.1111/ijsa.12314

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Shute, V. J., & Ventura, M. I. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press. http://library.oapen.org/handle/20.500.12657/26058

Smither, J. W., Reilly, R. R., Millsap, R. E., AT&T, K. P., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, *46*(1), 49–76. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Smits, J., & Charlier, N. (2011). Game-based assessment and the effect on test anxiety: A case study. In D. Gouscos, & M. Meimari (Eds.),

Proceedings of 5th European Conference on Games Based Learning (pp. 562–566). Academic Conferences International.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Truxillo, D. M., Bodner, T. E., Bertolino, M., Bauer, T. N., & Yonce, C. A. (2009). Effects of explanations on applicant reactions: A meta-analytic review. *International Journal of Selection and Assessment*, *17*(4), 346–361. https://doi.org/10.1111/j.1468-2389.2009.00478.x

Weidner, N., & Short, E. (2019). Playing with a purpose: The role of games and gamification in modern assessment practices. In R. N. Landers (Ed.), *The Cambridge handbook of technology and employee behavior* (pp. 151–178). Cambridge University Press. https://doi.org/10.1017/9781108649636.008

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. J. Sternberg, & J. Funke (Eds.), *The psychology of human thought* (pp. 227–247). Heidelberg University Publishing. https://doi.org/10.17885/heiup.470

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Wu, F. Y., Mulfinger, E., Alexander, L., Sinclair, A. L., McCloy, R. A., & Oswald, F. L. (2022). Individual differences at play: An investigation into measuring Big Five personality facets with game-based assessments. *International Journal of Selection and Assessment*, *30*(1), 62–81. https://doi.org/10.1111/ijsa.12360

How to cite this article: Ohlms, M. L., Melchers, K. G., & Kanning, U. P. (2024). Can we playfully measure cognitive ability? Construct-related validity and applicant reactions. *International Journal of Selection and Assessment*, *32*, 91–107. https://doi.org/10.1111/ijsa.12450