Received: 25 March 2022 | Accepted: 14 October 2022

DOI: 10.1111/bjet.13286

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Learning analytics application to examine validity and generalizability of game-based assessment for spatial reasoning

Yoon Jeon Kim[1] | **Mariah A. Knowles**[1] | **Jennifer Scianna**[1] | **Grace Lin**[2] | **José A. Ruipérez-Valiente**[3]

[1]Curriculum and Instruction, University of Wisconsin–Madison, Madison, Wisconsin, USA

[2]Education Arcade, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[3]Information and Communications Engineering, University of Murcia, Murcia, Spain

Correspondence

Yoon Jeon Kim, University of Wisconsin–Madison, Madison, WI 53715, USA.
Email: yj.kim@wisc.edu

Funding information

National Science Foundation

Abstract

Game-based assessment (GBA), a specific application of games for learning, has been recognized as an alternative form of assessment. While there is a substantive body of literature that supports the educational benefits of GBA, limited work investigates the validity and generalizability of such systems. In this paper, we describe applications of learning analytics methods to provide evidence for psychometric qualities of a digital GBA called Shadowspect, particularly to what extent Shadowspect is a robust assessment tool for middle school students' spatial reasoning skills. Our findings indicate that Shadowspect is a valid assessment for spatial reasoning skills, and it has comparable precision for both male and female students. In addition, students' enjoyment of the game is positively related to their overall competency as measured by the game regardless of the level of their existing spatial reasoning skills.

KEYWORDS

enjoyment, game-based assessment, generalizability, learning analytics, spatial reasoning, validity

INTRODUCTION

Games, both digital and non-digital, are increasingly seen as a considerable asset for learning (De Freitas, 2018; Gee, 2009; Prensky, 2003). This includes those games that have been

Practitioner notes

What is already known about this topic:

- Digital games can be a powerful context to support and assess student learning.
- Games as assessments need to meet certain psychometric qualities such as validity and generalizability.
- Learning analytics provide useful ways to establish assessment models for educational games, as well as to investigate their psychometric qualities.

What this paper adds:

- How a digital game can be coupled with learning analytics practices to assess spatial reasoning skills.
- How to evaluate psychometric qualities of game-based assessment using learning analytics techniques.
- Investigation of validity and generalizability of game-based assessment for spatial reasoning skills and the interplay of the game-based assessment with enjoyment.

Implications for practice and/or policy:

- Game-based assessments that incorporate learning analytics can be used as an alternative to pencil-and-paper tests to measure cognitive skills such as spatial reasoning.
- More training and assessment of spatial reasoning embedded in games can motivate students who might not be on the STEM tracks, thus broadening participation in STEM.
- Game-based learning and assessment researchers should consider possible factors that affect how certain populations of students enjoy educational games, so it does not further marginalize specific student populations.

explicitly designed for learning purposes, commonly called educational or serious games, by aligning the mechanics and content of the game with specific learning goals (Ibrahim & Jaafar, 2009), as well as games developed for entertainment since players can learn and refine multiple skills from active gameplay (Gee, 2009; Shute et al., 2015). Game-based assessment (GBA) is a specific use of educational games that employs game activities to elicit evidence for educationally valuable skills and knowledge (Kim et al., 2016).

Developing a well-balanced GBA, a game that is both enjoyable and holds satisfying qualities as assessment, is challenging (Kim et al., 2016), particularly related to ensuring and examining psychometric qualities such as validity, fairness and generalizability (Kim & Ifenthaler, 2019). While still limited, the field has begun to understand how students' backgrounds (eg, gender, game experience) differentially influence one's performance and learning in educational games (Hou et al., 2020) and validity and generalizability (eg, Auer et al., 2022). Because the underlying assumption for the benefits of games is that the player enjoys the game (Fu et al., 2009), to what extent the player enjoyed the game, and thus how much genuine effort the player put forth, should be considered for assessment (eg, Hou et al., 2020; Kim & Shute, 2015).

Moreover, GBA requires new ways of examining these psychometric qualities (Kim et al., 2016) because gameplay data can easily violate assumptions required for traditional statistical analysis (eg, an underlying latent variable stays static and has a normal distribution). While learning analytics and educational data mining methods could be used to overcome these challenges, the learning analytics community also recognizes the need to

strengthen the connection between learning analytics and assessment (Gašević et al., 2022) as little research has been done generally on properties of assessment in the field of learning analytics. Thus, examining these psychometric qualities is a key interest for the learning analytics and GBA communities, and the community has begun to apply interdisciplinary approaches to examine validity, generalizability and fairness in GBA, for example, using more commonly used methods in the learning analytics such as Bayesian Knowledge Tracing (Rizvi et al., 2022) and Random Forest (Auer et al., 2022).

In this paper, we report an evaluation of the validity and generalizability of GBA for spatial reasoning skills using a digital puzzle game called *Shadowspect*. The importance of this work to foster equitable and fair use of GBA in educational contexts can be established as follows. First, while there is a substantive body of literature that supports the importance of spatial skills for academic achievement and careers in STEM fields (eg, Uttal & Cohen, 2012), there are also well-documented gender differences in spatial reasoning skills (eg, Linn & Petersen, 1985) that could further mediate the gender difference in math and science (eg, Ganley et al., 2014; Nuttall et al., 2005). Second, playing video games has been shown to increase one's performance on spatial reasoning tasks for both males and females (eg, McClurg & Chaillé, 1987), games as a medium also have many known gender differences in terms of play styles and preferences, to name a few, males tend to prefer action and strategy, whereas females prefer social and physical, males tend to spend more time playing video games than females (eg, Hartmann & Klimmt, 2006; Scharkow et al., 2015). Therefore, to promote the increased use of games as a potential curriculum and assessment tool to support spatial reasoning in classrooms to mitigate gender inequalities in STEM learning, careful examination of validity and generalizability across male and female students should be one of the minimum requirements.

This paper investigates the following overarching question: 'Can a game be developed as an alternative assessment of spatial reasoning for curricular use?' We call attention throughout to evidence for the validity and generalizability of Shadowspect as an assessment tool for spatial reasoning skills using learning analytics techniques. Specifically, this paper addresses the following three research questions (RQs):

- **RQ1.** Validity: How well does Shadowspect assess spatial reasoning skills?
- **RQ2.** Generalizability: Does Shadowspect assess spatial reasoning skills with comparable reliability/precision for female versus male students?
- **RQ3.** Interplay with enjoyment: Does enjoyment affect the validity of Shadowspect?

RELEVANT LITERATURE

Game-based assessment

Well-designed games have the potential to provide a rigorous context for assessment and have been recognized and applied as an alternative to more traditional assessments (Shaffer & Gee, 2012) while providing the following affordances. First, games engage players in authentic environments through versatile mechanics where players adapt to specific rules and constraints, facilitating authentic situations that resemble what they will encounter in real-world situations (Bellotti et al., 2010). Second, the telemetry of digital games allows for collecting rich data that can be used to build computational models using psychometric and machine learning approaches. By doing so, we can reconstruct the entire problem-solving process instead of only looking at the final outcomes of the problem (Freire et al., 2016). Finally, annual surveys on youth media use consistently indicate that playing games is one of the most popular leisure activities (Anderson & Jiang, 2018), where 90 percent of teens

in the United States say they play video games of any kind (whether on a computer, game console, or cellphone). People's prior experience of playing games in their daily lives can lead to better engagement when introduced in educational contexts (Eyupoglu & Nietfeld, 2019). This positive attitude towards games allows increased data collection (due to more time spent on task) and increased overall accuracy of the inferences (due to lower test anxiety and higher engagement) (Courtney & Graham, 2019; Mavridis & Tsiatsos, 2017; Mulligan et al., 2018; Sundre & Wise, 2003; Wise, 2006).

The integration of player's background knowledge and prior experiences can create challenges for data collection as player in-game decisions are blended between prior and current understandings, thus creating noisier data (Basu et al., 2020). Yet, analysis of player data signatures using multiple event types can yield a more holistic understanding of player experience such as differentiating activity levels and proficiency (Pellicone et al., 2019). Data systems have continuously adjusted to include the scale of interactions that can be recorded from GBAs to make such distinctions by including not only player actions but also system events and player progress; this comprehensive log data allows for feature generation and ultimately better models (Owen & Baker, 2020). Recent work suggests that relatively few features may be needed to accurately predict student performance when they are meaningfully derived (Chen et al., 2020).

The current opportunities for GBA lie in the diversity of games which can be used as an assessment. GBA can come to be through an intentional design process where designers employ methods such as stealth assessment (Shute, 2011) or evidence-centred design (ECD) (Mislevy et al., 2003). However, even commercially available games can produce the types of interactional data necessary to assess player affect (Chen et al., 2020). From more concrete learning goals such as physics concepts (Kim et al., 2016) and computational thinking (Pellicone et al., 2019) to more abstract skills such as persistence (DiCerbo, 2014; Ventura & Shute, 2013), there are myriad ways to consider learner outcomes in GBA. In this work, we apply learning analytics to assess a more foundational mathematical skill, spatial reasoning.

Spatial reasoning

Spatial reasoning is a multidimensional cognitive ability that is often used interchangeably with other terms such as spatial thinking, spatial ability and visual thinking. Mulligan et al. (2018) provide a comprehensive definition for spatial reasoning as the 'ability to perform mental manipulations of visual stimuli, the ability to transform spatial forms into other visual arrangements, an awareness of the structural features of spatial forms and the analytical thinking required to find relationships and solve problems', (p. 78). Spatial reasoning not only is an essential skill to solve various problems in everyday life but also has been recognized as an important predictor for academic achievements and careers, in Science, Technology, Engineering and Math (STEM) and related fields including chemistry (Stieff, 2011), computer science, astronomy, physics, mechanical engineering, geometry and medicine (for a comprehensive review, see Uttal & Cohen, 2012).

Supporting spatial reasoning is of particular interest to reduce achievement gaps in STEM domains because it serves as a gateway skill to learning other academic disciplines where understanding abstract concepts in formal curriculum is often explained via visualizations (Uttal & Cohen, 2012); spatial reasoning is also associated with the development of early algebraic skills (Papic et al., 2011) and geometry performance (Clements & Battista, 1992). Many call for more training and assessment opportunities of spatial reasoning in schools (Wai & Uttal, 2018). For example, Wai and Uttal (2018) argue that the education system needs better curriculum and assessment tools for spatial reasoning to identify and support

talented students who might go unrecognized because the current assessment practices heavily rely on verbal reasoning and mathematics.

In addition, while there are many psychometrically validated spatial reasoning tasks, only a limited number of classroom-based spatial reasoning teaching and assessment tools exist (Ramful et al., 2017). The majority of the existing spatial reasoning instruments are pencil-and-paper where widespread use is limited; teachers often do not have adequate training to score and interpret test results or create appropriate interventions. Because many suggest that spatial reasoning skills are malleable and individuals can benefit from training opportunities, even with young adolescents, providing classroom-based tools that teachers can easily implement might lead to more equitable opportunities in developing spatial reasoning skills, therefore potentially decreasing the achievement gap in STEM learning (Lowrie & Jorgensen, 2018). In addition to the lack of resources, the existing research on spatial reasoning indicates the need for thoughtful approaches when it comes to individual differences, mainly student gender. That is, there has been a consistent report of gender differences between males and females, especially well-documented with mental rotation tasks. Law et al. (1993) additionally reported that the gender gap becomes wider with dynamic tasks. Therefore, we argue that understanding how fair or unfair GBA of spatial reasoning is across males and females is a crucial first step to suggesting more game-based curricular activities in classrooms to promote spatial reasoning skills regardless of gender.

Games to support and assess spatial reasoning

Games, both digital and non-digital, have been recognized as an effective training tool to improve people's spatial reasoning skills. For example, Olkun et al. (2005) created a digital format of Tangram puzzle game to train 4th and 5th graders' spatial visualization skills (ie, mentally manipulating pictures of objects to solve problems with visual/geometric content) and reported an increase in students' geometry score after playing the game for 80 to 120 minutes. Similarly, Yang and Chen (2010) had 5th graders play a digital format of a pentomino game for 60 minutes and reported a significant improvement of spatial reasoning skills between pre- and post-tests. While Yang and Chen reported that female students initially scored lower on the pre-test, the difference was decreased as the result of playing the game, suggesting that game-based interventions can serve to support certain subgroups. While there is sufficient evidence that supports the benefits of playing geometric games to train one's spatial reasoning skills, no work thus far has examined how games could function as an alternative assessment, nor what weaknesses and strengths GBA of spatial reasoning skills may have, particularly in a way that does not further reinforce the existing stereotype threats (eg, gamers, male).

Validity and generalizability in game-based assessment

The dictionary definition of psychometrics refers to the measurement of psychological attributes of individuals, but in practice, psychometrics refers ‘to a methodology to identify, characterize, synthesize and critique evidence in arguments about examinees' capabilities in light of the purpose and the context of assessment use’. (Mislevy et al., 2016). While the majority of existing psychometrics are based on the standard assessment paradigm as Mislevy et al. (2012) describe them as ‘discrete, pre-packaged tasks with just a few bits of data’ in the special issue of the Journal of Educational Data Mining, they call for new forms of complex assessment, such as GBA, to ‘embrace concepts and methods from educational data mining as well as from existing psychometrics’. That is, GBA designers should consider

psychometrics when designing the game and assessment models based on in-game data; proper consideration of psychometrics streamlines the connections between evidence generated in-game and the inferences that one wishes to make about learners (DiCerbo, 2014; Mislevy et al., 2016).

Like other forms of assessment, there are a set of qualities that GBA needs to demonstrate; we focus on validity and generalizability in this paper. In the following section, we discuss what these qualities are and how one can address these qualities in the context of GBA using interdisciplinary approaches including psychometrics and learning analytics methods.

The Standards for Educational and Psychological Testing defines validity as 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' (AERA, APA, & NCME, 2014, p. 9). Evidence for validity in GBA can be established in two stages: work done during the design phase of assessment and information gathered after the assessment is designed (DiCerbo et al., 2017). Many of the existing GBA works consider the first stage by explicitly using ECD as an assessment design framework (Mislevy et al., 2003). For example, for Shadowspect, the team designed coarse interactions within the game with target competencies in mind and generated features that are informed by the existing body of the literature, early design insights and experts' reviews (Kim et al., 2019). The present work considers the second stage, which seeks to align the prior interpretations with formal measures of assessment validity. Several existing GBA works consider the second stage using both psychometrics and learning analytics/educational data mining methods. For example, DiCerbo (2014) applied ECD to create two features (time and completion) related to persistence in an online game called Poptropica and conducted confirmatory factor analysis using scores based on the two features across three tasks. On the contrary, Chen et al. (2020) used a machine learning algorithm, that is, support vector machine with recursive feature elimination, to identify in-game features that have the most predictive power for the overall mastery in the game.

Generalizability can be understood as a general framework to 'estimate error variance associated with different sources of error' (AERA, APA, & NCME, 2014, p. 37) by investigating a number of factors that can significantly affect reliability/precision of the assessment. The Standards for Educational and Psychological Testing suggest three factors that could affect reliability/precision: assessment procedure, qualities of raters and the population (eg, one group's familiarity with assessment formats and instructions in the assessment procedure). Gender and race are two of the most commonly examined sociodemographic factors in relation to generalizability of assessment, and gender difference between male and female has been one of the biggest concerns for spatial reasoning assessment tools and has been extensively researched (Harris et al., 2021).

Furthermore, for Shadowspect, to what extent the overall accuracy of estimation one can make based on students' performance in the game across different subgroups is a particular concern, especially given that certain subgroups, for example, girls and students with high math anxiety, might be less comfortable with a GBA format that feels like math puzzles. Examining generalizability (ie, to what extent our model is reliable across genders) for Shadowspect is crucial because of the previously noted longstanding history of gender discrepancies on spatial reasoning assessments, including females scoring lower on prior instruments (Reilly et al., 2017).

Several existing GBA works investigated possible gender discrepancies in GBA using both psychometric and learning analytics and educational data mining approaches. For example, Kim & Shute, (2015) examined to what extent one's overall proficiency of conceptual physics understanding measured in a physics puzzle game called Physics Playground using the Pearson correlation coefficients between in-game performance measures and gender and gaming abilities. They report that the weights of evidence for the numbers of the

gold and silver badges (ie, features created for overall mastery in the game) vary between female and male players, indicating that these two features might not be generalizable. Hou et al. (2020) examined if male and female players differently benefit from playing a math game called Decimal Point using two versions of game learning- versus enjoyment-focused. They report players who are more enjoyment-focused learned more efficiently, and female players had higher learning gains than male players across all conditions. Similarly, Rizvi et al. (2022) investigated how female and male players learn differently in five language games in Navigo using Bayesian Knowledge Tracing (BKT) and reported that female players' overall learning rather was higher than male's. With these previous findings in mind, we turn our attention towards analysis of Shadowspect to consider whether it serves as a valid and generalizable GBA.

MATERIALS AND METHODS

Shadowspect

Shadowspect aims to explicitly measure common core Geometry standards (eg, visualize relationships between 2D and 3D objects) and spatial reasoning skills, while simultaneously measuring valuable attributes and behaviours in the game that are connected to lifelong outcomes, such as persistence. Our team consisted of an assessment scientist, a data scientist and a game designer; our development process leveraged the best practices of the three disciplines together (Kim et al., 2019). Unlike in the traditional ECD approach (Mislevy et al., 2003), we applied an iterative process with a rapid prototyping methodology that included multiple playtests, where for each of the playtests, we conducted a close analysis to ensure that we were maintaining the balance between game design, data modelling and assessment algorithms.

Figure 1 shows two puzzles with their respective correct solutions. When a puzzle begins, the player receives a set of silhouettes from different views that represent the figure the player needs to create by selecting and composing a set of primitive shapes. The primitive shapes the player can use are cubes, pyramids, ramps, cylinders, cones and spheres. Depending on the level and difficulty, the puzzle may restrict the quantity or type of shapes they can create. After putting these shapes in the scenario, they can also scale, move and rotate the shapes to build a figure that solves the puzzle. Students can move the camera to see the figure they are building from different perspectives, and they can use the 'snapshot' tool (ie, the camera icon on the right bottom corner) to generate a silhouette to see how close they are to the objective. Finally, the player can submit the solution, which the game will evaluate and provide feedback. A video of the gameplay is available online[1] and the game can also be played online.[2]

Data collection

The data used in this paper is collected from a classroom implementation of the game with two math teachers who teach four geometry classes at a public charter school in Massachusetts, USA. The teachers had the liberty to assign Shadowspect in whichever ways work with their curriculum, although the research team provided the recommendation of having the students attempt at least three intermediate and three advanced level puzzles. The teachers ended up implementing Shadowspect in their classrooms as homework during the 4-week unit on transforming geometric figures in Spring 2021. The study has been approved by the Institutional Review Board of a college in the northeastern United States. Because of

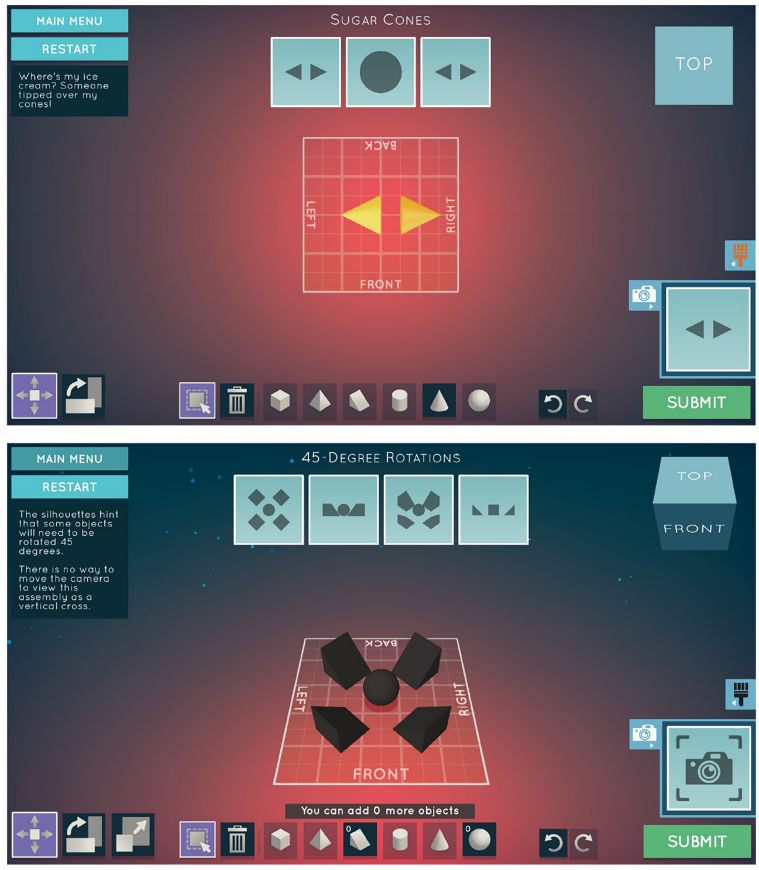

FIGURE 1 Two puzzle examples in Shadowspect

COVID-19 and the hybrid nature of classrooms, the consent and assent forms were sent electronically. To ensure student privacy, the research team did not collect student or parental email addresses; the collaborating classroom teacher sent out the consent forms to the parents instead. Additionally, students were asked to come up with their own nicknames when playing their game and to use the same nicknames on the various measures. Only the teacher had access to the student nicknames and their actual names. The research team never obtained students' names as an added measure of protection.

Two external measures of spatial reasoning were administered: Spatial Reasoning Instrument (SRI; [Ramful et al., 2017]) and Santa Barbara Solids Test, a cross-section test (SBST; [Cohen & Hegarty, 2012]). The SRI is a 30-item multiple choice instrument validated with middle school students with the internal reliability value of 0.849. Its internal consistency value was 0.84 in our study. SRI measures three sub-facets of spatial reasoning: mental rotation, spatial orientation and spatial visualization. While SRI measures the broad construct, SBST specifically focuses on one's ability to infer the two-dimensional cross-section of a three-dimensional object (ie, spatial visualization). SBST is a 30-item multiple choice test and has been validated with the target age group with a reliability of 0.91. Its internal consistency value was 0.89 in our study. Both assessments have a suggested completion time of 45 minutes. Because of the remote nature of schools due to COVID19, we converted the original paper-and-pencil tests into electronic versions using Qualtrics. The research team set a timer for 45 minutes (they would exit out of the survey after 45 minutes have passed).

We also administered a gameplay enjoyment questionnaire. Six of the items asked students about the gameplay experience using Shadowspect, a questionnaire adapted and validated from a validated instrument to measure one's motivation in games (Baker et al., 2006; Kim & Shute, 2015). In addition, we included 10 additional 5-point Likert scale items that have been validated and administered in the PISA assessment to assess students' math anxiety (eg, 'I get very nervous doing math problems') and self-efficacy (eg, 'I am not good at math') (Ferla et al., 2009).

From the four classes, we received full consent from students and parents to participate in the study from 61 students. A total of 44 students out of 61 completed the full set of spatial reasoning tests and questionnaires.

In addition to the external measures, Shadowspect also collected all the telemetry data that students generate while playing with the game. Any interaction with the game that students perform is stored as detailed data that allows us to reconstruct the learning process that students undergo to solve each puzzle. The game was developed as a Unity application, hosted as a web application, and all the events are emitted and stored in a MySQL database. We did not collect any identifiable or personal data from the users except for a nickname (login) the student provided themselves. We used that nickname login to merge the telemetry data from the game and the questionnaires.

The complete telemetry data collection from a total of the 44 students includes approximately 80,000 gameplay events (an average of 1818 events per user). Students were active in the game environment for a total of 52 hours (an average of 1.17 active hours per student), and students solved a total of 490 puzzles (an average of 11 puzzles per student).

Feature engineering

With the full data collection, we performed a feature engineering process to transform the raw telemetry data and external measure responses into the necessary features that would be used to respond the RQs of the study. From the telemetry data, we computed a number of features based on the functionality already implemented in Shadowspect for assessment purposes (Ruipérez-Valient et al., 2021). The implemented features are described in Table 1.

Model training and analyses

For RQ1, to investigate the validity of Shadowspect for spatial reasoning skills, we constructed Random Forest regression models (Breiman, 2001; Genuer et al., 2010) using *scikit-learn* to (i) measure the strength of the relationship between the suite of Shadowspect features (eg, persistence and competency scores) and the external measures of spatial reasoning and (ii) to identify the most important features the models. Two models were constructed, one each for our two external measures, SRI and SBST.

We chose Random Forest regression because it is well-suited to our needs. Once hyperparameters have been properly tuned, we eliminate the need for traditional feature selection necessary in other techniques, for example, linear regression (Genuer et al., 2010). This is desirable because we have several features (M = 30) and a small sample size (N_{sri} = 42, N_{sbst} = 41), owing to our data collection methods: several kinds of telemetry data were collected for two teacher's students. We tuned three hyperparameters that control the results of the algorithm: the number of trees in the forest, the max number of features per decision and max depth of any given tree (n_estimators, max features and max depth), following a grid search strategy (Genuer et al., 2010), once for each model.

TABLE 1 Student-level implemented features

Feature	Definition
tutorial_atime_comp	Seconds spent completing tutorial puzzles
total_breaks	Number of breaks in activity of more than 15 seconds
n_puzzle	Number of puzzles attempted (ignores repeats)
n_tutorial	Number of puzzles completed in the tutorial level
n_attempt	Number of attempts at puzzles (includes repeats)
n_complete	Number of puzzles successfully completed
revisits	Number of revisits to puzzles after failure
total_submit	Number of submitting a solution
incomplete_active	Active time for incomplete puzzles
complete_active	Active time for solved puzzles
tutorial_atime_complete	Active time spent for tutorials
n_failed_att	Total number of failed attempts
diff_event1	Total number of different events in tutorial level
avg_revisit	Average percent of revisits after failing
p-puz_no_basic	Percent of non-tutorial puzzles that were attempted
total_snapshot	Number of snapshot events
total_manipulate	Number of events manipulating a shape
total_view	Number of events from changing the camera angle
different_event	Number of different events
cumu_a_time	Time on task as a percentile
comp_average	The average competency score is computed based on a multivariate Elo-based learner modelling algorithm, using the four Common Core Standards at the level of 6–9th grades (Ruipérez-Valient et al., 2022) present in Shadowspect puzzles, individually represented by the features comp_mg1, comp_gmd4, comp_co5 and comp_co6
persistence_score	The persistence score is computed based on three dimensions: the time, the number of attempts, and the number of events when solving each puzzle. The metric takes into account the percentile distribution for each one of these dimensions, meaning that students are categorized as persistence or not compared with other peers in the class
non_per	Percent of time spent non-persistent in an unsolved puzzle
productive	Percent of time spent persisting in a puzzle and solving it
unproductive	Percent of time spent persisting in an unsolved puzzle
rapid	Percent of time spent without need to persist in a puzzle because the puzzle was solved much faster than peers
no_beh	Percent of time spent without any persistence type above

After tuning, models were *k*-fold cross-validated ($k = 5$) using a stratified sample to ensure that gender ratios within random folds resembled the overall observed gender ratio as closely as possible. The results of the models were predictions for each student for each external measure based on what was predicted when that student was in the test set ($\hat{Y}_{sri}$ and $\hat{Y}_{sbst}$). These predictions were only used to understand the strength of the relationships between our suite of features and our external measures, for which we computed the Pearson correlations (r_{sri} and r_{sbst}) between these model predictions and observed ground truths (y_{sri} and

y_{sbst}). Finally, to check the face validity of these results, we computed the GINI importance, a drop-out loss method, for each feature (Loecher, 2020).

For RQ2, to demonstrate comparable reliability across different subgroups, we computed the Pearson correlation between model predictions and observed ground truths of only the male students ($r^{(b)}_{sri}$ and $r^{(b)}_{sbst}$) and of only the female students ($r^{(g)}_{sri}$ and $r^{(g)}_{sbst}$). We tested the significance of the difference between these results using a two-tailed test with Fisher-*z* transform.

For RQ3, to demonstrate the effect of enjoyment on key features identified in analysis of RQ1, we ran a set of nested model ANOVAs, predicting the features of interest given student's self-reported enjoyment of Shadowspect, holding constant their observed spatial reasoning skills (SRI and SBST). The key features of interest were time on task (*cumu_a_time*), persistence (*persistence_score*) and competency score (*comp_average*).

RESULTS

RQ1

The overall performance of the two Random Forest regression models were comparable, although the model for SRI performed slightly better (Table 2). The correlation between SRI predicted and observed was r_{sri} = 0.54, whereas the correlation between SBST predicted and observed was r_{sbst} = 0.31. Under normal social science contexts, these would be considered moderate and weak correlations respectively; however, within the area of GBA we could consider these strong and moderate values respectively. The rationale is that transforming game data into constructs is quite a challenging process that often does not lead to high-performing results, and these values are quite good for the GBA field (Kim & Ifenthaler, 2019).

The most important features (Table 3) of both models included time on task (cumu_a_time), some measure of competency (comp_a_time) and some measure of persistence (persistence score or rapid). In contrast, the least important features of both models included counts related to puzzles attempted (eg, n_attempt and n_puzzle) and counts related to specific in-game events (eg, total_snapshot and total_manipulate). Together, it appears that both feature-engineered variables and scores from the learner modelling algorithms that Shadowspect uses were more important than raw telemetry data.

Inspecting the partial dependence (Figure 2) of the features associated with time on task, competency and persistence, in both models, holding all else as observed, shows that on average, (i) predictions of spatial reasoning increase as students receive higher competency scores, (ii) predictions increase as students score higher on the *rapid* persistence measure, (iii) predictions generally decrease as students spend more time on task (ie, *cumu_a_time*) and (iv) predictions decrease as students score higher on the overall persistence score. For example, as *cumu_ a_time* increases beyond 60, the mean predicted value of SRI, holding all else as observed, sharply decreases by over 10 points, indicating a cut score around 60 in the RF model and a negative relationship between *cumu_a_time* and SRI.

TABLE 2 Random forest regression performance

Model	*N*	*r*	R^2	MAE (% of error)	RMSE (% of error)
SRI	42	54%	28%	4.4 (15%)	5.0 (17%)
SBST	41	31%	0.2%	5.7 (19%)	7.0 (23%)

TABLE 3 Most and least important features

SRI (3.643)	SBST (2.244)
cumu_a_time (3.939)	rapid (3.085)
comp_average (3.861)	comp_mg1 (3.021)
comp_co5 (3.837)	cumu_a_time (2.944)
comp_gmd4 (3.810)	comp_co5 (2.929)
persistence_score (3.806)	comp_average (2.751)
comp_co6 (3.795)	n_tutorial (2.700)
…	…
n_attempt (3.658)	n_failed_att (2.327)
complete_active (3.657)	revisits (2.317)
total_snapshot (3.653	complete_active (2.314)
p_puz_no_basic (3.653)	n_puzzle (2.303)
non_per (3.651)	total_manipulate (2.292)
revisits (3.646)	n_attempt (2.279)

RQ2

On average, male students scored higher on both external measures ($\mu_{sri}^{(b)} = 19.5, \mu_{sri}^{(g)} = 14.83, \mu_{sbst}^{(b)} = 14.67, \mu_{sbst}^{(g)} = 11.65$). These differences were significant for both measures ($df_{sri} = 42$, $F_{sri} = 1.777$, $p_{sri} = 0.1896$, $df_{sbst} = 39$, $F_{sbst} = 1.834$, $p_{sbst} = 0.1835$). Considering the better of the two models (SRI), the model performance at predicting this external measure was slightly better for female than male students (Table 4). However, the difference between correlations ($r_{sri}^{(b)}$ and $r_{sri}^{(g)}$) was not significant ($z = 0.919$, $p = 0.179$). Given the limits of the sample size, it appears, cautiously, that Shadowspect has comparable reliability/precision for both female and male students.

RQ3

Students' self-reported enjoyment of playing Shadowspect had a positive relationship with their overall competency as measured by the game (*comp_average*), even when controlling for either external measure of prior spatial reasoning (Table 5). On average, each 1-point increase in enjoyment (one level higher across a set of 5-point Likert scales) corresponded with an increase of about 0.06 in competency score (six percentage points higher across a set of percentile measures). However, knowing a student's enjoyment of the game does no better at understanding the student's persistence (*persistence_score*) or time on task (*cumu_a_time*) than simply controlling for either external measure.

DISCUSSION

In this paper, we apply learning analytics techniques to evaluate the psychometric qualities of a GBA called Shadowspect, providing evidence for its validity and generalizability. In addition, we examined the extent one's enjoyment with the game affects the validity of Shadowspect as an assessment. Our findings suggest that we have sufficient evidence for the validity of Shadowspect as an assessment of spatial reasoning skills for middle school students. In-game, feature–based, Random Forest models used to predict two external

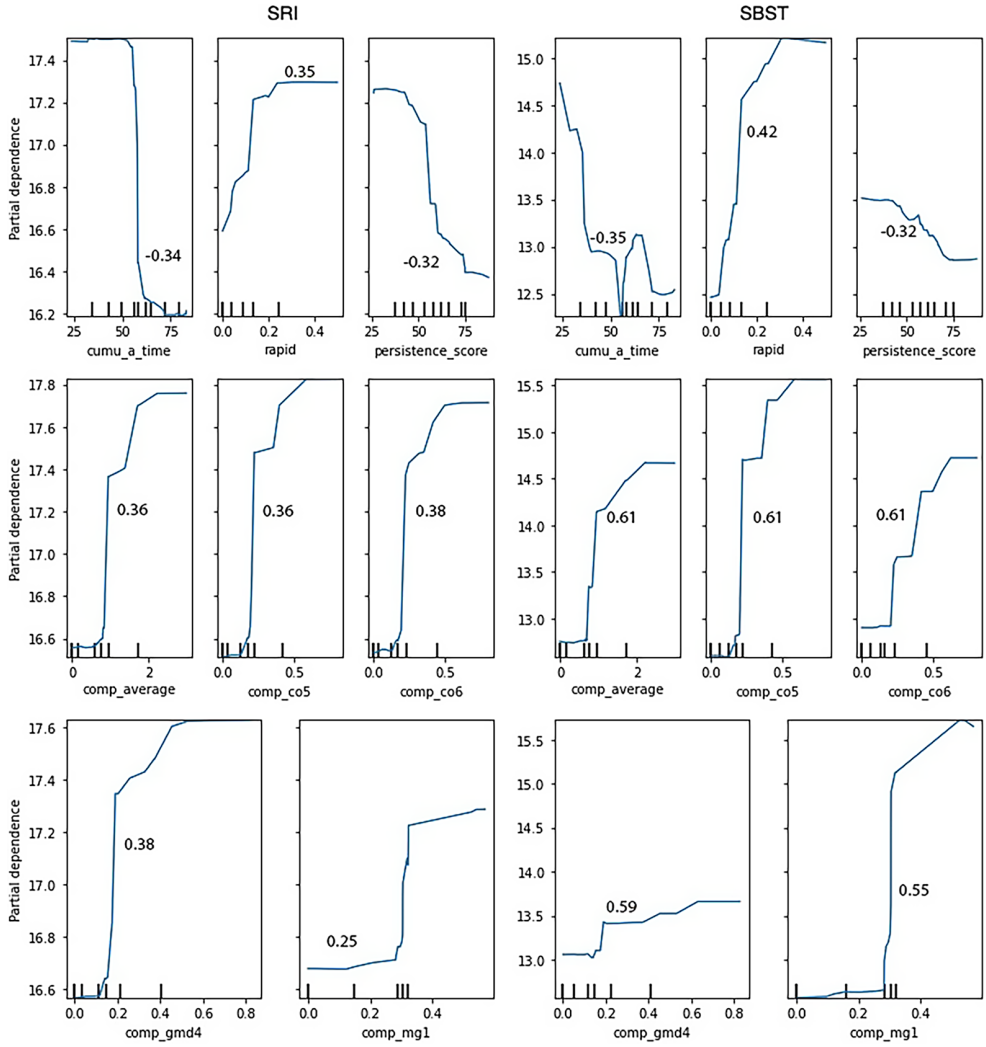

FIGURE 2 Partial dependence plots of top features, with pearson correlation coefficients

TABLE 4 SRI model performance by gender

Gender	*N*	*R*	R^2	MAE (% of error)	*RMSE* (% of error)
Male	18	34%	2.6%	5.3 (18%)	5.6 (19%)
Female	24	58%	27%	3.7 (12%)	4.4 (15%)

measures, SRI and SBST, have good model performances indicating that Shadowspect not only gets at the overall spatial reasoning skill but also a sub-facet, spatial visualization. Regarding the most influential features, the RF models for SRI and SBST have similar sets of variables. Both models included scores from the learner modelling algorithms that Shadowspect uses (ie, Elo scores and persistence) and features that were engineered rather than raw telemetry data. Our findings also indicate that Shadowspect as a whole has comparable reliability/precision for both male and female students, although the evidence is quite weak. Given the gender gap has been a persisting concern for spatial reasoning tests, further

TABLE 5 Enjoyment-related ANOVA Results. Nested model ANOVAs predicting *comp_average* (a, b), predicting *persistence_score* (c, d), *predicting cumu_a_time* (e, f), controlling for the SRI external measure (a, c, e) and controlling for the SBST external measure (b, d, f)

	a_0	a_1	b_0	b_1	c_0	c_1
Intercept	−0.0576	0.0308	−0.2939	−0.2946	68.2956	68.3527
External measure	0.0394	0.0307	0.0637	0.0595	−0.7164	−0.7220
Enjoyment	…	0.0632	…	0.0672	…	0.0408
R^2	0.126	0.222	0.370	0.493	0.105	0.105
F	…	5.028**	…	9.410***	…	0.00458
	d_0	d_1	e_0	e_1	f_0	f_1
Intercept	65.9487	65.9506	69.8038	69.9303	67.4475	67.4489
External measure	−0.6868	−0.6756	−0.7788	−0.7913	−0.7771	−0.7691
Enjoyment	…	−0.1775	…	0.0904	…	−0.1265
R^2	0.102	0.105	0.116	0.116	0.121	0.122
F	…	0.08856	…	0.02124	…	0.04252

**$p < 0.05$; ***$p < 0.01$.

investigation with a larger sample size is needed to fully understand how the set of features used to create spatial reasoning models can be generalizable.

Lastly, our findings indicate that the validity and generalizability of the game can be influenced by the extent one enjoyed playing the game. Our nested ANOVA models indicate that one's enjoyment significantly affects one key feature, *comp_average*, but not features associated with time on task or persistence. This finding is consistent with what others have reported regarding how one's motivation and engagement with the task affect the validity of the assessment (Sundre & Wise, 2003; Wise, 2006). In our work, we caution this in terms of fairness, especially related to equal opportunities to learn in the classroom. That is, teachers who are using Shadowspect in classrooms should account for some students simply not enjoying playing the game, and therefore, any suggestions from the assessment for 'low proficiency' should be interpreted with this caveat in mind. This is especially relevant given that female students' disadvantage associated with games and spatial reasoning tasks is known, thus teachers need to consider to what extent all students, not just male students and self-identifying gamers, enjoy playing the chosen game and how the teacher can accommodate those students who simply might not enjoy playing the game.

The present study provides a few implications for game-based learning and assessment. First, empirically investigating the psychometric qualities of the game-based learning system using learning analytics techniques can provide an opportunity for researchers and developers to understand how the students interacted with the game and how valid and reasonable one's inferences about students learning in the game may be. Second, while the underlying premise of game-based learning and assessment is that young people enjoy playing games and have positive associations with games, we should not assume that all students will enjoy playing games, especially when used in classroom settings. Therefore, game-based learning and assessment researchers need to carefully examine how certain students might be ultimately disadvantaged from the game-based system. Similarly, these qualities of the GBA need to be clearly communicated to the teachers, so teachers can provide appropriate pedagogical accommodations for those students. Furthermore, while these issues were investigated in the context of games in the present work, our findings can be broadly

applicable for similar EdTech genres such as VR and Simulations as these technologies also rely on students' problem-solving processes in highly interactive and data-rich environments.

This study has a few limitations. First, Shadowspect is currently not developed in a way that students who need visual or auditory accommodations can play. Therefore, our sample excludes students with such disabilities. Second, because of the difficulty getting students consented, we ended up with a data set of only 44 students with an unbalanced number of males and females. Future work with a larger sample size can provide stronger evidence regarding the validity and generalizability of Shadowspect as an assessment of spatial reasoning. Third, while the gender difference is a well-documented concern in the literature of spatial reasoning and game-based learning, the current study did not investigate how it plays out in Shadowspect, nor how teachers can use Shadowpect in classrooms to support female students. Future studies can further investigate how the teachers who implement the game in classrooms need to consider these psychometric qualities of Shadowspect to not further penalize female students. For example, teachers can ensure that female students are spending sufficient time playing the tutorial and easy puzzles until they are fully ready to move onto more challenging puzzles. Teachers also could have pairs of male and female students collaboratively play a few puzzles first and debrief different strategies that the pairs used. Third, without proper teacher training regarding how these algorithms work and how female students might need additional support, there is a real danger of reinforcing teachers' existing biases.

CONCLUSION

GBA gained popularity over the past decade, and its use has been increasing for various purposes, from formative use in the classroom to high-stakes hiring decisions. Like any other form of assessment, GBA needs to provide evidence for its psychometric qualities to support valid use in practice. Learning analytics practices offer a new way to investigate such qualities and are particularly useful for aggregating what the game intends to assess using numerous clickstream data. Our work further supports what the fields of learning analytics and GBA have been claiming, that we need an interdisciplinary approach to assessment in complex, data-rich, technology-enhanced learning environments such as digital games. Furthermore, this work is the first to investigate how games can be used to assess spatial reasoning skills, particularly related to the gender differences that have been well documented in the previous literature.

FUNDING INFORMATION

This work was supported by National Science Foundation Grant Number 1935450.

CONFLICT OF INTEREST

The authors have no conflict of interest.

DATA AVAILABILITY STATEMENT

We made both our data and Python scripts publicly available on our GitHub repository: https://github.com/EducationalSciences693/ShadowspectRandomForest.

ETHICS STATEMENT

This material is the authors' own original work, which has not been previously published elsewhere.

ORCID

Yoon Jeon Kim https://orcid.org/0000-0002-3467-1238
Mariah A. Knowles https://orcid.org/0000-0002-1798-4830
Jennifer Scianna https://orcid.org/0000-0003-1029-3452
Grace Lin https://orcid.org/0000-0001-7552-2880
José A. Ruipérez-Valiente https://orcid.org/0000-0002-2304-6365

ENDNOTES

[1] Video trailer of Shadowspect: https://youtu.be/j1w_bOvFNzM.

[2] Playable version of Shadowspect: https://fielddaylab.wisc.edu/play/shadowspect/.

REFERENCES

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing: National Council on Measurement in Education*. American Educational Research Association.

Anderson, M., & Jiang, J., Placeholder Text (2018). Teens, social media & technology 2018. *Pew Research Center*, *31*(2018), & 1673–1689.

Auer, E. M., Mersy, G., Marin, S., Blaik, J., & Landers, R. N. (2022). Using machine learning to model trace behavioral data from a game-based assessment. *International Journal of Selection and Assessment*, *30*(1), 82–102.

Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems* (Vol. 2002, pp. 29–36).

Basu, S., Disalvo, B., Rutstein, D., Xu, Y., Roschelle, J., & Holbert, N. (2020). The role of evidence centered design and participatory design in a playful assessment for computational thinking about data. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 985–991). The Association for Computing Machinery.

Bellotti, F., Berta, R., & De Gloria, A. (2010). Designing effective serious games: Opportunities and challenges for research. *International Journal of Emerging Technologies in Learning*, *5*(2010), 22–35.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, *30*(3), 481–503.

Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. In D. A. Grouws(Ed.), *Handbook of research on mathematics teaching and learning* (pp. 420–464). Macmillan.

Cohen, C. A., & Hegarty, M. (2012). Inferring cross sections of 3d objects: A new spatial thinking test. *Learning and Individual Differences*, *22*(6), 868–874.

Courtney, L., & Graham, S. (2019). 'It's like having a test but in a fun way' young learners' perceptions of a digital game-based assessment of early language learning. *Language Teaching for Young Learners*, *1*, 161–186.

De Freitas, S. (2018). Are games effective learning tools? a review of educational games. *Journal of Educational Technology & Society*, *21*(2), 74–84.

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

DiCerbo, K. E., Shute, V., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–21). Springer.

Eyupoglu, T. F., & Nietfeld, J. L. (2019). Intrinsic motivation in game-based learning environments. In *Game-based assessment revisited*(pp. 85–102). Springer.

Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic selfconcept: Reconsidering structural relationships. *Learning and Individual Differences*, *19*(4), 499–505.

Freire, M., Serrano-Laguna, Á., Manero, B., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. In *Learning, design, and technology* (pp. 1–29). Springer Nature.

Fu, F. L., Su, R. C., & Yu, S. C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, *52*(1), 101–112.

Ganley, C. M., Vasilyeva, M., & Dulaney, A. (2014). Spatial ability mediates the gender difference in middle school students' science performance. *Child Development*, *85*(4), 1419–1432.

Gašević, D., Greiff, S., & Shaffer, D. W. (2022). Towards Strengthening Links between Learning Analytics and Assessment: Challenges and Potentials of a Promising New Bond. *Computers in Human Behavior*, 107304.

Gee, J. P. (2009). Video games, learning, and "content". In *Games: Purpose and potential in education* (pp. 43–53). Springer.
Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.
Harris, D., Lowrie, T., Logan, T., & Hegarty, M. (2021). Spatial reasoning, mathematics, and gender: Do spatial constructs differ in their contribution to performance?*British Journal of Educational Psychology*, *91*(1), 409–441.
Hartmann, T., & Klimmt, C. (2006). Gender and computer games: Exploring females' dislikes. *Journal of Computer-Mediated Communication*, *11*(4), 910–931.
Hou, X., Nguyen, H. A., Richey, J. E., & McLaren, B. M. (2020). Exploring how gender and enjoyment impact learning in a digital learning game. In *International Conference on Artificial Intelligence in Education* (pp. 255–268). Springer, Cham.
Ibrahim, R., & Jaafar, A. (2009). Educational games (eg) design framework: Combination of game design, pedagogy and content modeling. In *2009 international conference on electrical engineering and informatics* (Vol. 1, pp. 293–298). IEEE
Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, *16*(2), 142–163.
Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In *Game-based assessment revisited* (pp. 3–11). Springer.
Kim, Y. J., Ruipérez-Valiente, J. A., Tan, P., Rosenheck, L., & Klopfer, E. (2019). Towards a process to integrate learning analytics and evidence-centered design for game-based assessment. In *Companion Proceedings of the 9th International Learning Analytics and Knowledge Conference* (pp. 204–205). The Association for Computing Machinery.
Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, *87*, 340–356.
Law, D. J., Pellegrino, J. W., & Hunt, E. B. (1993). Comparing the tortoise and the hare: Gender differences and experience in dynamic spatial reasoning tasks. *Psychological Science*, *4*(1), 35–40.
Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*(6), 1479–1498.
Loecher, M. (2020). Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods*, *51*, 1–13.
Lowrie, T., & Jorgensen, R. (2018). Equity and spatial reasoning: Reducing the mathematical achievement gap in gender and social disadvantage. *Mathematics Education Research Journal*, *30*(1), 65–75.
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.
McClurg, P. A., & Chaillé, C. (1987). Computer games: Environments for developing spatial cognition? *Journal of Educational Computing Research*, *3*(1), 95–111.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), 1–29.
Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.
Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., vonDavier, A., & John, M. (2016). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 23–48). Routledge.
Mulligan, J., Woolcott, G., Mitchelmore, M., & Davis, B. (2018). Connecting mathematics learning through spatial reasoning. *Mathematics Education Research Journal*, *30*(1), 77–87.
Nuttall, R. L., Casey, M. B., & Pezaris, E. (2005). Spatial ability as a mediator of gender differences on mathematics tests: A biological-environmental framework. In A. M. Gallagher & J. C. Kaufman(Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 121–142). Cambridge University Press.
Olkun, S., Altun, A., & Smith, G. (2005). Computers and 2d geometric learning of Turkish fourth and fifth graders. *British Journal of Educational Technology*, *36*(2), 317–326.
Owen, V. E., & Baker, R. S. (2020). Fueling prediction of player decisions: Foundations of feature engineering for optimized behavior modeling in serious games. *Technology, Knowledge and Learning*, *25*(2), 225–250.
Papic, M. M., Mulligan, J. T., & Mitchelmore, M. C. (2011). Assessing the development of preschoolers' mathematical patterning. *Journal for Research in Mathematics Education*, *42*(3), 237–268.
Pellicone, A., Holbert, N., Disalvo, B., Kumar, V., & Berland, M. (2019). Who played the game correctly? data signatures of interaction in playful assessment. In *Proceedings of the 2019 connected learning summit* (pp. 138–146).
Prensky, M. (2003). Digital game-based learning. *Computers in Entertainment (CIE)*, *1*(1), 21.
Ramful, A., Lowrie, T., & Logan, T. (2017). Measurement of spatial ability: Construction and validation of the spatial reasoning instrument for middle school students. *Journal of Psychoeducational Assessment*, *35*(7), 709–727.

Reilly, D., Neumann, D. L., & Andrews, G. (2017). Gender differences in spatial ability: Implications for stem education and approaches to reducing the gender gap for parents and educators. In *Visual-spatial ability in stem education* (pp. 195–224). Springer.

Rizvi, S., Gauthier, A., Cukurova, M., & Mavrikis, M. (2022). Examining Gender Differences in Game-Based Learning Through BKT Parameter Estimation. In *International Conference on Artificial Intelligence in Education* (pp. 600–606). Springer, Cham.

Ruipérez-Valient, J. A., Gomez, M. J., Martínez, P. A., & Kim, Y. J. (2021). Ideating and developing a visualization dashboard to support teachers using educational games in the classroom. *IEEE Access*, *9*, 83467–83481.

Ruipérez-Valient, J. A., Kim, Y. J., Baker, R. S., Martínez, P. A., & Lin, G. C. (2022). The affordances of multivariate elo-based learner modeling in game-based assessment. *IEEE Transactions on Learning Technologies*. 1–14. https://ieeexplore.ieee.org/abstract/document/9875051

Scharkow, M., Festl, R., Vogelgesang, J., & Quandt, T. (2015). Beyond the "core-gamer": Genre preferences and gratifications in computer games. *Computers in Human Behavior*, *44*, 293–298.

Shaffer, D. W., & Gee, J. P. (2012). The right kind of gate: Computer games and the future of assessment. In *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 211–228). Information Age Publishing, Inc.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, *80*, 58–67.

Stieff, M. (2011). When is a molecule three dimensional? a task-specific role for imagistic reasoning in advanced chemistry. *Science Education*, *95*(2), 310–336.

Sundre, D. L., & Wise, S. L. (2003). Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Uttal, D., & Cohen, C. (2012). Spatial thinking and STEM education: When, why and how. In *Psychology of learning and motivation* (Vol. 57, pp. 147–181). Academic Press.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572.

Wai, J., & Uttal, D. H. (2018). *Why spatial reasoning matters for education policy*. American Enterprise Institute.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114.

Yang, J. C., & Chen, S. Y. (2010, November). Effects of gender differences and spatial abilities within a digital pentominoes game. *Computers & Education*, *55*(3), 1220–1233.

How to cite this article: Kim, Y.J., Knowles, M.A., Scianna, J., Lin, G., & Ruipérez-Valiente, J.A. (2023). Learning analytics application to examine validity and generalizability of game-based assessment for spatial reasoning. *British Journal of Educational Technology*, *54*, 355–372. https://doi.org/10.1111/bjet.13286