Computers in Human Behavior 119 (2021) 106701

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Construction and validation of a game-based intelligence assessment in minecraft

Heinrich Peters [a,*], Andrew Kyngdon [b], David Stillwell [c,d]

[a] Columbia Business School, Columbia University, USA
[b] New South Wales Education Standards Authority, USA
[c] The Psychometrics Centre, University of Cambridge, USA
[d] Judge Business School, University of Cambridge, USA

ARTICLE INFO

Keywords:
Game-based assessment
Intelligence assessment
Minecraft
Project malmo
Process data
Game log-data

ABSTRACT

Video games are a promising tool for the psychometric assessment of cognitive abilities. They can present novel task types and answer formats, they can record process data, and they can be highly motivating for test takers. This paper introduces the first game-based intelligence assessment implemented in *Minecraft*, an exceptionally popular video game with more than 200 m copies sold. A matrix-based pattern completion task (PC), a mental rotation task (MR) and a spatial construction task (SC) were implemented in the three-dimensional, immersive environment of the game. PC was intended as a measure of inductive reasoning, whereas MR and SC were measures of spatial ability. We tested 129 children aged 10–12 years old on the *Minecraft*-based tests as well as equivalent pen-and-paper tests. All three scales fit the Rasch model and were moderately reliable. Factorial validity was good with regard to the distinction between PC and SC, but no distinct factor was found for MR. Convergent validity was good as abilities measured with *Minecraft* and conventional tests were highly correlated at the latent level ($r = 0.72$). Subtest-level correlations were in the moderate range. Furthermore, we found that behavioral log-data collected from the game environment was highly predictive of performance in the *Minecraft* test and, to a lesser extent, also predicted scores in conventional tests. We identify a number of behavioral features associated with spatial reasoning ability, demonstrating the utility of analyzing granular behavioral data in addition to traditional response formats. Overall, our findings indicate that *Minecraft* is a suitable platform for game-based intelligence assessment and encourage future work aiming to explore game-based problem solving tasks that would not be feasible on paper or in conventional computer-based tests.

1. Introduction

1.1. *Game-based intelligence assessment*

Video games involve cognitive processes and performance in certain video games is positively correlated with intellectual ability (Quiroga et al., 2009, 2011). Early approaches to measuring cognitive performance through video games date back to the 1980s with Mané and Donchin’s (1989) space fortress game (see also Jones, 1984; Jones, Dunlap, & Bilodeau, 1986; Rabbitt, Banerji, & Szymanski, 1989). More recent work suggests that intellectual ability can be reliably captured with commercial video games (Foroughi, Serraino, Parasuraman, & Boehm-Davis, 2016; Quiroga et al., 2015, Quiroga, Román, De La Fuente, Privado, & Colom, 2016). As these examples indicate, the term game-based assessment describes the use of (video) games for the measurement of psychological constructs. As such, it is related but distinct from concepts like gamification (”the use of game design elements in non-game contexts”; Deterding, Dixon, Khaled, & Nacke, 2011), serious games (the use of games for the purpose of learning and education; Deterding et al., 2011; Gee, 2003), computerized assessment (the use of computer interfaces to present testing materials), and simulations or complex problem solving tasks (using interactive virtual environments to present dynamic tasks that mimic real-world problems; Dorner & Funke, 2017; Greiff & Funke, 2009; Greiff, Wüstenberg, Molnár, Fischer, Funke, & Csapó, 2013).

The construction of intelligence assessments is deeply intertwined with a long history of research into the structure of intelligence. Intelligence is widely understood as a multi-faceted construct and several

* Corresponding author.
E-mail addresses: hp2500@columbia.edu (H. Peters), Andrew.Kyngdon@nesa.nsw.edu.au (A. Kyngdon), d.stillwell@jbs.cam.ac.uk (D. Stillwell).

https://doi.org/10.1016/j.chb.2021.106701
Received 22 January 2020; Received in revised form 28 December 2020; Accepted 10 January 2021
Available online 19 January 2021
0747-5632/© 2021 Elsevier Ltd. All rights reserved.

prominent models of the structure of intelligence have been proposed in the literature (e.g., Carroll, 1993; Horn, 1968; Horn & Blankson, 2005; Horn & Cattell, 1966). The most comprehensive and most widely accepted model to date is the Cattell-Horn-Carroll (CHC) model, which integrates the works of Raymond Cattell, John Horn, and John Carroll (Flanagan & Dixon, 2014; Schneider & McGrew, 2012). Like Carroll's (1993) Three-Stratum Theory, the CHC hierarchically distinguishes between "narrow abilities" (stratum I), "broad abilities" (stratum II) and a single "general ability" or g-factor (stratum III). The broad abilities specified in the CHC provide a particularly useful taxonomy of higher level cognitive abilities. Among others, they include Fluid Intelligence (Gf), Crystallized Intelligence (Gc), General Knowledge (Gkn), Quantitative Knowledge (Gq), Reading/Writing Ability (Grw), Short-Term Memory (Gsm), Long-Term Storage and Retrieval (Glr), Visual Processing (Gv), Auditory Processing (Ga), and Processing Speed (Gs). Most new and recently revised intelligence test batteries are based on the taxonomy suggested in the CHC model (Flanagan & Dixon, 2014; Schneider & McGrew, 2012).

Compared to conventional intelligence assessment with its centennial history, game-based intelligence assessment is still in its fledgling phase and none of the existing game-based intelligence assessments captures the full granularity of the CHC model. Nonetheless, game-based intelligence assessment offers a promising perspective as it combines the flexibility of simulations and the motivational benefits of games and gamification. It therefore has the potential to address two important limitations of conventional intelligence assessments, namely their reliance on static, two-dimensional testing materials (Foroughi et al., 2016; Hunt, 2011; Jodoin, 2003) and their limited ability to motivate test takers (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011).

The possibilities of psychometric test design are greatly enhanced by computerization and the use of video games, which enable complex, interactive stimuli that are impracticable in conventional assessment. These new possibilities can be leveraged to either address limitations of existing measures or to assess psychological functions that could previously not be captured for practical reasons (Hunt & Pellegrino, 1985). For example, game-based presentation facilitates the measurement of abilities associated with spatial reasoning, such as navigation or reasoning about moving objects (Jackson, Vernon, & Jackson, 1993). The same is true for time-sensitive task formats as used for the measurements of attention (Godwin, Lomas, Koedinger, & Fisher, 2015) and dynamic task types, where the test taker has to explore and interact with the testing materials like in complex problem solving tasks (Greiff, Niepel, Scherer, & Martin, 2016). Additionally, video games allow for scoring approaches incorporating process data, such as action sequences or movement profiles, along with traditional formats focusing on outcome data, where the psychometric information of an item is typically restricted to a right-or-wrong dichotomy. This is especially advantageous when the process leading to the solution contains information about a person's ability level see (Bergner, Shu, & von Davier, 2014; Hao, Shu, & von Davier, 2015; Shu, Bergner, Zhu, Hao, & von Davier, 2017; Zhu, Shu, & von Davier, 2016).

A fundamental problem of intelligence assessment is the susceptibility of test results to non-ability-related factors, such as motivation (Borghans, Meijers, & ter Weel, 2013; Duckworth et al., 2011) or test anxiety (Meijer & Oostdam, 2007; Oostdam & Meijer, 2003). A meta-analysis by Duckworth et al. (2011) suggests that test-taking motivation accounts not only for a large proportion of the variance in test results, but also for a large proportion of the variance in academic achievement and other life outcomes that have otherwise often been attributed to intelligence. The authors showed that incentives can increase IQ test results on average by 0.64 *SD*, with the highest gains in below-average IQ participants, indicating that IQ tests do not always measure maximum intellectual performance and that low results may be due to motivational deficits. Furthermore, the study found that the predictive validity of IQ scores dropped significantly after adjusting for the influence of test motivation, indicating that test motivation is a third variable inflating the predictive power of IQ scores. Borghans et al. (2013) found that intrinsic motivation (enjoyment of the task) as well as extrinsic motivation (presence of a monetary incentive) improved performance on IQ-tests by increasing participants' time investments per item. With regard to test anxiety, Meijer and Oostdam (2007) found consistent negative relationships between levels of anxiety and results in verbal ability, reasoning and memory tests. Game-based assessments, however, have been associated with higher levels of motivation in testing situations (Lumsden, Edwards, Lawrence, Coyle, & Munafò, 2016) and lower degrees of test anxiety (Mavridis & Tsiatsos, 2017).

Despite their potential advantages, game-based intelligence assessments are rare. This is explained by a range of practical and theoretical challenges concerning the development of new task types, administration modes, and psychometric models for process data and complex answer formats. First, as Washburn (2003) points out, the development of game-based tasks requires more effort and expertise compared to conventional tasks, as it often involves the creation of computer applications from scratch. Second, the high degree of flexibility that is characteristic of computer games places special emphasis on the trade-off between standardization and generalization, and traditional psychometric concepts like the assumption of unidimensionality become less sustainable. Third, there may be conflicts between principles of game design and psychometric task design. Feedback, for example, is a central motivational aspect of games (Przybylski, Rigby, & Ryan, 2010), but can create dependencies between items, which is not desired according to the traditional psychometric paradigm. Fourth, there is the question of how to treat confounding variables specific to game-based assessment, such as psychomotor factors and familiarity with computer technology or games (Foroughi et al., 2016; Greiff, Kretzschmar, Müller, Spinath, & Martin, 2014; Washburn, 2003). These points are essential to the questions of fairness and validity, as design decisions may have differential effects on different groups of test takers (Mislevy et al., 2014), and may change the construct that is being assessed at the latent level (Hunt & Pellegrino, 1985; Mead & Drasgow, 1993; Quiroga et al., 2016). Finally, there are cultural challenges, referring to the strong tradition of originally paper-based intelligence assessments and their ascendancy in psychological research and practice, a point that is repeatedly criticized by Hunt (2011). The issue is illustrated by the fact that intelligence is still mostly measured with pen-and-paper tests, even when computerized versions of the same test are psychometrically equivalent and favored by test takers (Arce-Ferrer & Martínez Guzmán, 2009; Quiroga et al., 2016).

While the research community has shown interest in game-based assessment of cognitive abilities since the 1980s (Haier, Siegel, Tang, Abel, & Buchsbaum, 1992; Jones, 1984; Jones et al., 1986; Mané & Donchin, 1989; Rabbitt, Banerji, & Szymanski, 1989), recent advances in computer and entertainment technology have reinvigorated this field of research. Previous work can be roughly divided into two lines of research, one of which is concerned with the creation of custom games designed for the sole purpose of measurement (Mané & Donchin, 1989; McPherson & Burns, 2008; McPherson & Burns, 2007; Ventura, Shute, Wright, & Zhao, 2013), while the other is concerned with the exploration of the psychometric properties of existing, commercially available video games (Baniqued et al., 2013; Buford and O'Leary, 2015; Foroughi et al., 2016; Haier, Siegel, Tang, Abel, & Buchsbaum, 1992; Jones et al., 1986; Quiroga et al., 2015; Quiroga et al., 2016; ; Quiroga, Diaz, Román, Privado, & Colom, 2019). Representing the first of these two lines of research, McPherson & Burns developed the games *Space Code* (McPherson & Burns, 2007) and *Space Matrix* (McPherson & Burns, 2008), where test takers had to solve tasks resembling the WAIS-III Digit Symbol subtest and the Dot Matrix task (Miyake et al., 2000) in order to destroy enemy space-ships. Correlations between *Space Code* scores and Gs ranged between 0.45 and 0.60, while correlations between *Space Matrix* and Gy ranged from 0.53 to 0.66. Ventura et al. (2013) developed a new measure of spatial ability, involving a navigation task where test-takers have to use a first-person avatar to collect gems in a range of virtual spaces. Performance in the task was significantly correlated with three other measures of spatial ability, standardized math test scores and STEM career choices. However, self-reported experience with video games was also associated with higher test scores, indicating that game experience may confound performance scores.

The second line of game-based assessment research is concerned with the use of existing video games for intelligence measurement. Baniqued, Lee, Voss, Basak, Cosman, DeSouza, Severson, Salthouse, & Kramer (2013) analyzed relationships between performance in 20 web-based video games and a battery of cognitive tasks, finding correlations in the moderate range. Quiroga et al. (2015) showed that participants' performance in a series of puzzles from the video game *Big Brain Academy* is strongly correlated with a wide range of conventional intelligence tasks at the latent level ($r = 0.93$). *Big Brain Academy* is a collection of puzzle games intended to train and measure a range of cognitive abilities, which are similar to the known factors of intelligence. In two subsequent studies, Quiroga et al. (2016) made use of another puzzle game, *Professor Layton and the Curious Village*, to investigate the relationship between game performance over time and intelligence. Correlations between game performance and intelligence measures peaked after about seven to 12 hours of game play. Completion time was inversely correlated with intelligence measures, suggesting that mastery of a complex task is indicative of cognitive ability. While the study provides important insights into the relationships between game performance and intelligence, the game seems too long to be used as an assessment in practice in its current form. Foroughi et al. (2016) created a customized game-based assessment with the *Portal 2 Puzzle Creator*. *Portal 2* is a first-person puzzle-platform game, where the player has to solve problems and manipulate a three-dimensional environment in order to escape from a series of chambers. Performance in the game-based test is highly correlated with Ravens Advanced Progressive Matrices ($r = 0.65$) and a latent factor representing fluid intelligence ($r = 0.78$). Similar results have been reported by Buford & O'Leary (2015). More recently, Quiroga et al. (2019) analyzed the relationships between performance in a wide range of commercially available video games and the stratum II intelligence factors fluid reasoning (Gf), visuo-spatial ability (Gv), and processing speed (Gs). The results confirm previous findings, showing a stong latent correlation between factors representing game performance and intelligence ($r = 0.79$).

While these studies suggest that game-based assessments may be psychometrically valid, they did not investigate process data from game log files, as commercial video games typically limit access to game-logs and do not provide opportunities to modify the underlying code (Quiroga et al., 2016). The use of process data was recently investigated in the *Wells* task (National assessment of educational progress, 2014), a complex problem solving task, in which test takers have to repair a virtual mechanic water pump. Hao et al. (2015) found that edit distances between students' action sequences and the action sequence of the optimal solution are strongly correlated with overall scores, indicating that the edit distance to the best performance sequence contains information about test takers' ability levels. Representing action sequences as weighted directed networks, Zhu et al. (2016) showed that network measures such as weighted density, centralization and reciprocity predict students' scores on the task. Bergner et al. (2014) performed cluster analyses on action sequences and found relationships between cluster membership and scores, however, varying greatly with the clustering method that was employed. Shu et al. (2017) propose a new Item Response Theory (IRT) model to derive ability estimates from process data. This is achieved by modelling action sequences as a Markov process and treating its state transitions as items in an IRT model, where the probability of choosing a transition depends on the latent trait. Methods like these and others (e.g., Fu, Zapata, & Mavronikolas, 2014) could also be used to tap into process data for game-based intelligence assessment.

1.2. Contributions of the present study

Previous work has shown that the use of existing video games is a promising option for the creation of game-based assessments. However, as Quiroga et al. (2016) point out, existing games are usually inflexible with regard to task design and data export. As a consequence, none of the game-based assessments reviewed above has explored the psychometric properties of process data captured from the game environment. For the present study, we used the popular video game *Minecraft* in combination with *Project Malmo* (Johnson, Hofmann, Hutton, & Bignell, 2016), an application programming interface (API) providing full control over the game environment, to implement a series of visuospatial and inductive reasoning tests. We were thus able to leverage the game environment of an immensely popular video game while retaining full control over task design and capturing high-resolution log-data for further analysis. The present work combines the approaches of Baniqued et al. (2013), Buford and O'Leary (2015), Foroughi et al. (2016) and Quiroga et al. (2015, 2016, 2019), who pioneered the use commercial video games for intelligence assessment, and Bergner et al. (2014), Hao et al. (2015), Zhu et al. (2016) and Shu et al. (2017), who have started to explore the psychometric properties of process data in complex problem solving.

Minecraft is a digital sandbox game that enables players to design and simulate three-dimensional worlds. The player is represented as a first-person avatar in a voxel-based environment consisting of cubic blocks. Players can create fanciful structures from a variety of materials, interact with other players and accept quests such as solving puzzles or fighting monsters. Producing raw materials through mining activities and collecting items in order to build structures are central aspects of the game. With more than 200 million copies sold since its release in 2009, *Minecraft* is among the most popular games of all time. Its popularity as well as its flexibility have recently led *Minecraft* to being used in educational applications and research (Dikkers, 2015; Ekaputra, Lim, & Eng, 2013; Ellison, Evans, & Pike, 2016; Karsenti, Bugmann, & Gros, 2017; Pusey & Pusey, 2015). *Minecraft* has also shown potential as a tool for artificial intelligence (AI) research. *Project Malmo* is an AI experimentation platform based on *Minecraft*. It contains an API that enables researchers to integrate AI agents, design tasks and run experiments in the game environment (Johnson et al., 2016). The API provides full control over the configuration of the *Minecraft* environment, including spatial arrangements, target structures and data export.

We used *Project Malmo* to create a game-based assessment in *Minecraft*. The game incorporates a narrative, extensive tutorials and currently three *Minecraft*-based intelligence tasks: a matrix-based pattern completion task (PC) where the player has to infer a set of rules in order to complete a series of matrices, a mental rotation task (MR), and a spatial construction task (SC), similar to traditional block design tasks, where the player has to recreate a model structure with a limited inventory of blocks. The tasks can be classified according to the Cattell-Horn-Carrol model (Carroll, 1993; McGrew, 2005). The PC task would be classified as a measure of inductive reasoning belonging to the factor of fluid intelligence (Gf). The MR and SC tasks involve visuo-spatial abilities and would thus be categorized as visual processing tasks (Gv). However, the *Minecraft*-based tasks differ from their conventional counterparts in several important ways: they are implemented in an immersive three-dimensional environment that can be explored and manipulated; they focus heavily on constructed response formats; they record process data in addition to outcome data; and they are integrated in a game structure with a narrative. Hence, our *Minecraft*-based testing platform incorporates many of the advantages of game-based assessments and provides a solution to most of the practical concerns discussed in the previous section, as the *Project Malmo* API provides full control over the game environment, enabling the construction of customized task types, stimuli and data export functions in *Minecraft*. The design of the *Minecraft*-based testing platform was guided by the recommendations made by Quiroga et al. (2011, 2015, 2016). A detailed description of the testing platform and the implemented tasks can be found in the methods section.

The main goal of this paper is to investigate *Minecraft* and *Project Malmo* as a tool for game-based intelligence assessment. We first conduct Rasch analyses at subscale level in order to eliminate psychometrically poor items and to assess the psychometric properties of the three scales. Secondly, we assess factorial validity. We hypothesize that the items of the three scales load on distinct factors representing distinct but positively related abilities. Third, convergent validity is assessed by analyzing the relationship between the subscales of the *Minecraft*-based test and two conventional reasoning tests: Raven's Standard Progressive Matrices (RSPM; Raven, Raven, & Court, 2000) and Vandenberg & Kuse Mental Rotations Test (VKMR; Peters et al., 1995). In line with previous research (Foroughi et al., 2016; Quiroga et al., 2015), we hypothesize a strong positive relationship between latent factors of intellectual ability as measured in *Minecraft* and measured by conventional means. With regard to discriminant validity, we expect *Minecraft*-based subscales to show weaker relationships with paper-based tests belonging to a different factor in the Cattell-Horn-Carrol model (Carroll, 1993; McGrew, 2005) as compared to correlations with constructs of the same factor. Furthermore, we analyze effects of gender and prior experience with the video game *Minecraft* on ability scores. Here, we hypothesize a positive effect of Minecraft experience on performance in the SC task, as the task requires relatively good command of the game controls. We also expect that there is a gender effect regarding performance in the MR and SC tasks, consistent with prior research indicating a pronounced male advantage in spatial reasoning (Voyer, Voyer, & Bryden, 1995). Finally, we explore the relationships between game log-data from the SC task and performance in all *Minecraft*-based tasks, as well as the conventional validation tests. First, we use supervised machine learning techniques to predict scale scores from the log-data on a hold-out set not previously used for modelling. We expect the SC log-data to be most predictive of performance in the SC task itself as compared to performance in the PC and the MR tasks if the log-data captures information about task-specific abilities. Similarly, we expect the SC log-data to be more predictive of VKMR scores as compared to RSPM scores if the log-data captures specific information about visuo-spatial reasoning (Gv). Second, we use unsupervised learning to find clusters in the log-data, which represent distinct behavioral patterns. We expect the different behavioral clusters to be associated with different levels of performance in the SC task. We use local regression models to further analyze the nonlinear relationships between individual behavioral features and performance.

2. Method

2.1. Participants

The three *Minecraft*-based tests as well as the validation tests Raven's Standard Progressive Matrices (RSPM) and Vandenberg & Kuse Mental Rotations Test (VKMR) were administered to a sample of Australian fifth and sixth grade primary school students. A total of 130 students took the *Minecraft*-based test. The age of the participants ranged from ten to twelve years ($M = 11.18$, $SD = 0.73$). In the sample there were 73 girls and 57 boys from six different classes in two schools. Out of the 130 participants, 116 had played the video game *Minecraft* before. A total of 120 data sets could be matched with paper-based test results from the RSPM and VKMR tests (69 girls, 106 with *Minecraft* experience; age $M = 11.18$, $SD = 0.73$). The unmatched data was still used for scale construction and factor analyses. All other analyses were performed on the subset of data that could be matched with paper-based test results. No individual participants were excluded from the analyses. The schools were public primary schools in metropolitan Sydney chosen by the New South Wales Education Standards Authority (NESA). School principals selected which classes of grade 5 or grade 6 students participated.

2.2. Materials

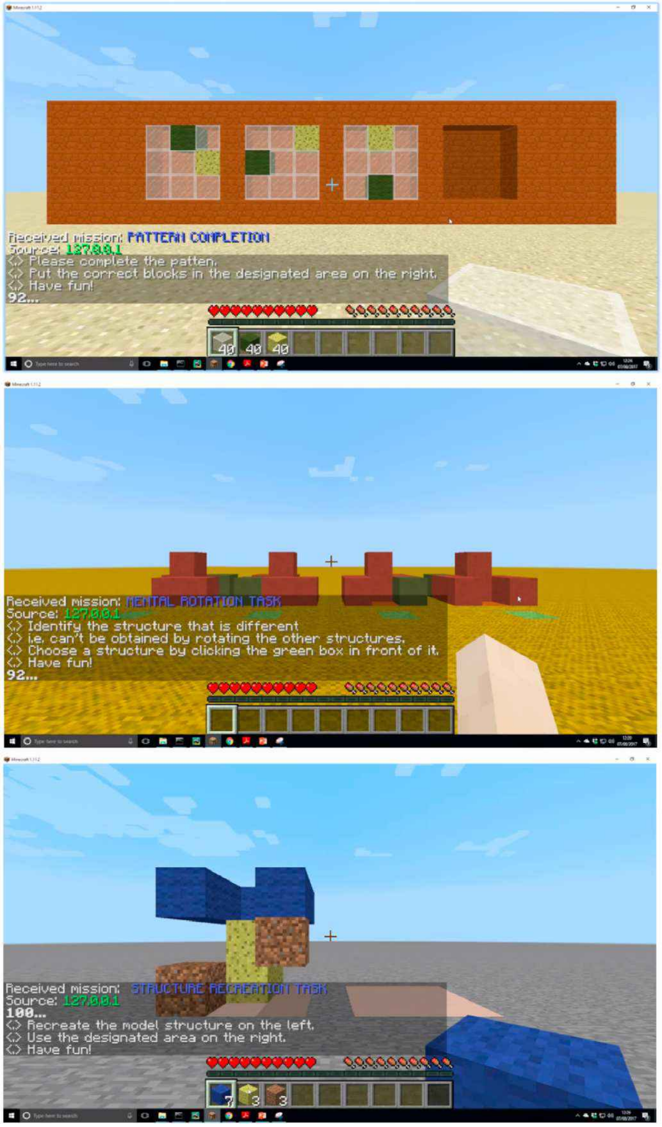
The *Minecraft*-based test was created by the authors of the present paper using the *Project Malmo* API (Johnson et al., 2016) for the programming language Python (van Rossum, 1995). A key feature of the platform is its ability to record log data from the *Minecraft* game environment through a customized data export function. The version that was used in the present study comprised the three intelligence tasks Pattern Completion (PC), Mental Rotation (MR) and Spatial Construction (SC). Each subtest consisted of 12 items, some of which were later excluded as a result of the item analysis. Tutorials were used to familiarize test takers with the game controls and to minimize the effect of prior *Minecraft* experience. The first tutorial orients the test taker to basic keyboard and mouse controls. The second tutorial demonstrates how to destroy blocks and collect raw materials. The third tutorial trains the test taker to use their inventory and to build structures using the materials mined in the previous tutorial. At the end of the tutorials, there was a practice test requiring test takers to demonstrate basic command of the game controls by placing colored blocks in designated areas. The tutorial section was followed by the PC task, the MR task and the SC task, all of which are described in detail in the measures section. In order to succeed in the tasks, no controls were needed that had not been previously covered in the tutorials. All other controls that exist in the commercially available version of *Minecraft* were disabled. Tutorials, questionnaires and subtests were integrated in a space-themed narrative which we assumed would create a coherent game-like experience.

2.3. Measures

The measures in the present study fall into four broad categories. First, there are the actual *Minecraft*-based scales PC, MR and SC. Second, there are other performance related measures from the *Minecraft* game environment, such as log data and data from the tutorial tasks. Third, there are self-report measures including demographics and control variables. And finally, there are the paper-based validation tests Raven's Standard Progressive Matrices (RSPM; Raven, Raven, & Court, 2000) and Vandenberg & Kuse Mental Rotations Test (VKMR; Peters et al., 1995). The two validation tests were chosen, because they are well established measures of fluid intelligence and spacial reasoning, suitable for administration in group settings.

Pattern completion task (PC). The first *Minecraft*-based subtest is a matrix-based inductive reasoning test, where the test taker completes a sequence of matrices according to a set of underlying transformations, which have to be inferred from visual information. Each item presents a series of three 3 × 3 matrices consisting of two to four different block colors arranged in a varying pattern. The matrices form a horizontal row, embedded in a wall facing the avatar. The matrix on the left specifies the initial configuration of each series. All other matrices in a specific item are transformations of the matrix to their left, which means they can be obtained by applying a set of rules to the blocks in the preceding matrix. Blocks of the same type are subject to identical rules: they move around and appear or disappear in the same recurring pattern. The fourth matrix in each series is a blank space that can be filled with blocks from the avatar's inventory. If the test taker fills in the correct blocks in the correct pattern, the answer is scored as correct. Test takers are allowed to delete and replace blocks. Each item has a time limit of 100 s. If the test taker does not solve the item within 100 s, the solution is scored as incorrect. The PC task is intended as a measure of fluid intelligence. The scale was constructed such that items with higher complexity, in terms of the number of different block types and the number of rules they involve, were expected to be more difficult. An example of a PC item can be found in Fig. 1.

Mental rotation task (MR). The second *Minecraft*-based subtest is a mental rotation task. In each item the test taker encounters a range of four structures in three-dimensional space. Each structure is placed on a 3 × 3 plane and reaches up to three levels in height. Three out of the four



Received mission: PATTERN COMPLETION
Source: 127.0.0.1
<.> Please complete the patten.
<.> Put the correct blocks in the designated area on the right.
<.> Have fun!
92...

Received mission: MENTAL ROTATION TASK
Source: 127.0.0.1
<.> Identify the structure that is different
<.> i.e. can't be obtained by rotating the other structures.
<.> Choose a structure by clicking the green box in front of it.
<.> Have fun!
92...

Received mission: STRUCTURE RECREATION TASK
Source: 127.0.0.1
100...
<.> Recreate the model structure on the left.
<.> Use the designated area on the right.
<.> Have fun!

Fig. 1. Screenshot of an of *Minecraft*-based Pattern Completion (PC) item, Mental Rotation (MR) item, and a Spatial Construction (SC) item in this order. Video examples can be found at https://github.com/hp2500/MARS.

structures in each set are identical but rotated along at least one rotation axis. One out of the four structures is different, i.e. it cannot be obtained by rotating any of the other three structures. The task requires the test taker to determine which of the structures is different. The answer is submitted by clicking on a green square that is embedded in the ground in front of each structure. Choosing the correct structure is scored as a correct answer. Choosing the wrong structure or exceeding the time limit of 100 s per item is scored as an incorrect answer. A multiple choice answer format was chosen, because constructed response instructions would have been hard to comprehend by the participants (i.e. we would have needed to specify for each item by how many degrees, along which of the three rotation axes, and in which directions the structure should be rotated). As such, the *Minecraft*-based MR task is very similar to its pen-and-paper equivalent, the Vandenberg & Kuse Mental Rotations Test (Peters et al., 1995). A crucial difference is that test takers can actively change their position relative to the model structures. Being able to see the structures from different angles puts emphasis on spatial exploration as a means of gathering information that is necessary to solve the task. The scale was constructed such that items with a larger number of blocks, larger degrees of rotation and more rotation axes were expected to be more difficult. An example of an MR item can be found in Fig. 1.

Spatial construction task (SC). The third *Minecraft*-based subtest can be described as a virtual block design test. The test taker has to copy a three-dimensional model structure to a designated area in virtual space. Each model is placed on a 3 × 3 plane and reaches up to three levels in height. The test taker can move the avatar around to investigate the model structure. The avatar is equipped with a minimal inventory of blocks, just enough to recreate the model structure. As a result the test taker is sometimes forced to plan ahead and build temporary scaffolds to place blocks that could otherwise not be placed, e.g. levitating blocks. If the structure in the designated building area is identical to the model structure, the answer is scored as correct. Each item has a time limit of 100 s. If the test taker does not solve the item within 100 s, the solution is scored as incorrect. The SC task is intended as a measure of spatial ability. The scale was constructed such that items with more blocks and higher complexity in terms of scaffolding were expected to be more difficult. An example of a SC item can be found in Fig. 1.

Tutorial Test. The tutorial test requires participants to demonstrate basic command of the game controls which they should have learned from the tutorials, i.e. navigating through their inventory and manipulating their environment by placing blocks. The tutorial test is a single item where the participant has to place four differently colored blocks on four designated areas of matching colors. Time spent on the tutorial test is used as a measure of proficiency with the game controls. The test has a time limit of 100 s.

Self-report measures. As control variables we collected age, gender (m/f) and prior *Minecraft* experience (y/n) as well as positive attitude towards video games and self-assessed gaming skills on a five-point Likert scale (see SI 1). Additionally, a test enjoyment questionnaire (TEQ) was designed to capture how much the participants liked the *Minecraft*-based test (see SI 2).

Log data. Log data was recorded from the game environment in real time with a resolution of up to 20 times per second. The log data can be classified into time, space and action-related variables. Time-related variables include response times, time spent on tutorials, instructions and narrative screens as well as timestamps for every other recorded piece of data. Spatial data includes an avatar's position in space (x, y, z coordinates), orientation in space (yaw, pitch) and detailed ray-casting information (coordinates of fixated point, type of fixated block, distance to fixated block). Action data captures the test taker's interactions with relevant parts of the environment, namely correctly and incorrectly placed blocks as well as the number of corrections. Essentially, enough data was collected to completely replay an individual participant's gameplay. Several aggregate-level variables were also recorded. These included distance travelled, distance between the avatar and relevant structures in the environment, and how long the cursor was pointed at relevant parts of the environment, such as answer options, or specific structures that were part of the task.

2.4. Design and procedure

The tests were administered in classroom settings of about 25 students per session as part of the students' normal school day. The *Minecraft*-based tests and validation tests were administered in two separate blocks of less than 1 h each, divided by a 30-min lunch break. The two blocks were counterbalanced so that half of the sample worked through the *Minecraft*-based tests first, while the other half of the sample worked through the pen-and-paper tests first. Both administrations were invigilated and the participants were made aware of the fact that they were in a testing situation. Laptops were supplied to students with the *Minecraft* test application pre-installed. No personal or school laptops were used.

At the beginning of the *Minecraft*-test block the participants were prompted to fill out a consent form and complete an introduction screen asking for demographics and other personal information, namely their attitude towards video games, self-assessed gaming skills and *Minecraft* experience (see SI 1). After completing the initial questionnaire, participants worked through the three tutorials and the tutorial test. In all three tutorials, the test takers were required to finish the task within a 100-s time limit. If a test taker failed a tutorial task, the tutorial was reset and the test taker had to try again until the task was successfully completed. This was to ensure that all test takers understood the controls well enough to succeed in the actual subtests and to minimize the effect of prior experience with the *Minecraft* game controls. After the tutorials, the PC task, the MR task and the SC task were completed in this order. The first two items of each task were training items. At the end of the *Minecraft*-test block, the participants completed the test enjoyment questionnaire (see SI 2).

In the pen-and-paper block the participants first completed the VKMR and then the RSPM. The VKMR was administered first, because it is a timed test, which had the advantage that all participants finished the test roughly at the same time. The RSPM test was administered with a time limit of 40 min. Participants who finished early were instructed to stay at their desks and remain quiet.

3. Results

3.1. Rasch analyses and scale construction

The item analyses of the three *Minecraft*-based subscales were based on the Rasch model. The Rasch model is an item-response theory (IRT) model, in which the probability of a specific item response (in our case correct/incorrect) is modeled as a logistic function of a person parameter (ability) and an item parameter (difficulty). Higher difficulty of an item is associated with a lower probability for test takers to solve the item correctly. Test takers with higher ability, on the other hand, are more likely to solve items correctly. The parameters are fitted such that, given a specific difficulty parameter value for each item and a specific ability parameter value for each test-taker, the overall likelihood of the observed data is maximized (Smith & Smith, 2004). In order to create scales that were in line with the assumptions of the Rasch model, we used stepwise item elimination. The first two items of every subscale were training items and were therefore excluded from the start. Other items were sequentially excluded from each subscale in case of low variance (more than 95% of the cases in one answer category) and significant misfit ($p < .05$) in the χ^2-based item total fit statistic (e.g., Smith & Smith, 2004). In each iteration the item with the highest χ^2-value was eliminated until no item significantly misfitted the model. Exceptions to this rule were made in cases where the exclusion of such items would have resulted in a problematic overall model fit in Andersen's likelihood ratio test or scale length dropping below six items. Andersen's likelihood ratio test allows to assess whether parameters estimates differ between subsamples created by spliting the overall sample. If the model holds, the parameter estimates do not vary significantly between the subsamples (Engelhard, 2013). All steps of the analysis as well as parameter estimation were performed with the R package eRm (Mair, Hatzinger, & Maier, 2009).

In the three *Minecraft*-based subtests PC, MR and SC, 4.9%, 5.5% and 3.1% of the data was missing due to technical glitches during test administration. The data was missing completely at random (MCAR) according to Little's (1988) MCAR test (all $p > .054$). The values were imputed using a random forest algorithm (Stekhoven & Buhlmann, 2012), a non-parametric technique that has proven to be effective for imputation of responses in the context of Rasch analysis (Golino and Gomes, 2016).

In the PC scale, item pc4 was excluded due to low variance and item pc12 due to low item fit. Andersen's likelihood ratio tests demonstrated acceptable overall model fit under the median split criterion ($\chi^2(7) = 10.59$, $p = .158$) and the mean split criterion ($\chi^2(7) = 7.401$, $p = .388$). In the final 8-item scale, part-whole corrected point-biserial correlations ranged from 0.28 to 0.57 ($M = 0.43$, $SD = 0.10$). Cronbach's alpha was $\alpha = 0.735$. In the MR scale, items mr9, mr12 and mr11 were excluded due to low item fit. Item mr7 was retained despite low item fit, as the item was essential to overall model fit. Overall model fit in Andersen's likelihood ratio test was acceptable under the mean split criterion ($\chi^2(6) = 4.28$, $p = .64$) and under the median split criterion ($\chi^2(6) = 2.67$ ($p = .85$). Part-whole corrected point-biserial correlations in the final 7-item scale ranged from 0.21 to 0.42 ($M = 0.33$, $SD = 0.069$). Cronbach's alpha was $\alpha = 0.61$. In the SC scale, items sc3 and sc5 were excluded due to low variance and items sc11 and sc4 were excluded due to low item fit. Item sc9 was retained despite marginal item misfit, as it correlated well with the rest of the scale and scale length was not to be further reduced. Andersen's likelihood ratio tests were not meaningful due to a lack of variance in some of the subgroups produced by the splits. Alternative measures of model fit, specifically Pearson's R^2 and area under the receiver-operating-characteristic curve (AUC) as proposed by Mair et al. (2008), demonstrated good model fit ($R^2 = 0.69$, $AUC = 0.96$). In the final 6-item scale, part-whole corrected point-biserial correlations ranged from 0.35 to 0.61 ($M = 0.50$, $SD = 0.10$). Cronbach's alpha was $\alpha = 0.76$. Item level statistics for all *Minecraft*-based subscales can be found in Table 1. For scale level statistics, please see Table 2. For a graphical depiction of item information curves, see SI 3. Additional psychometric analyses based on classical test theory (CTT) are reported in SI 4.

3.2. Construct validity

3.2.1. Factorial validity

We tested for factorial validity across all three subscales. For this purpose we conducted a confirmatory factor analysis using the lavaan R package (Rosseel, 2012). We assumed a three-factor structure with latent variables corresponding to the three *Minecraft*-based scales and a single higher order factor representing g. Individual items of the reduced scales were used as indicators. PC items were expected to load on one factor representing inductive reasoning and MR and SC items were expected to load on two factors representing distinct spatial abilities. The estimation method was maximum likelihood and we used standardized

Table 1
Results of item analyses - retained items.

item	diff	se	cor	χ^2-fit	df	p	sp
PC							
pc3	−1.44	0.24	0.43	83.82	114	0.99	0.81
pc9	−0.77	0.21	0.48	79.84	114	1.00	0.71
pc10	−0.72	0.21	0.28	127.61	114	0.22	0.71
pc5	−0.67	0.21	0.57	66.92	114	1.00	0.70
pc8	0.00	0.20	0.41	123.52	114	0.30	0.59
pc7	0.55	0.20	0.39	122.14	114	0.33	0.50
pc6	0.68	0.20	0.54	91.74	114	0.95	0.47
pc11	2.37	0.26	0.32	108.32	114	0.68	0.21
MR							
mr5	−0.35	0.18	0.36	99.03	110	0.76	0.49
mr3	−0.31	0.18	0.29	118.59	110	0.27	0.48
mr4	−0.27	0.18	0.29	115.67	110	0.34	0.47
mr7	−0.20	0.18	0.21	139.25	110	0.03	0.46
mr6	0.08	0.18	0.35	104.64	110	0.63	0.40
mr8	0.08	0.18	0.42	95.77	110	0.83	0.40
mr10	0.97	0.21	0.37	98.13	110	0.78	0.25
SC							
sc6	−3.26	0.37	0.42	47.60	108	1.00	0.84
sc7	−2.34	0.32	0.52	123.58	108	0.14	0.75
sc8	−1.14	0.29	0.61	66.80	108	1.00	0.62
sc9	1.36	0.29	0.53	141.86	108	0.02	0.34
sc10	1.56	0.29	0.58	44.49	108	1.00	0.32
sc12	3.83	0.43	0.35	29.82	108	1.00	0.10

Note: PC = pattern completion task, MR = mental rotation task, SC = spatial construction task, diff = item difficulty, se = standard error, cor = part-whole corrected point-biserial correlations of item and scale, χ^2 = chi-square item total fit, df = degrees of freedom of item fit statistic, p = significance level of item fit statistic, sp = proportion of participants who solved the item.

Table 2
Scale level statistics.

	M	*SD*	miss	LR	AUC	BIC	Alpha
PC	0.55	1.73	4.89%	.39	.87	616	.74
MR	−0.40	1.42	5.55%	.638	.77	663	.61
SC	−0.06	2.70	3.10%	/	.96	217	.76
VKMR	12.0	5.73	/	/	/	/	.90
RSPM	33.7	4.92	/	/	/	/	.76

Note. PC = pattern completion task, MR = mental rotation task, SC = spatial construction task, VKMR = Vandenberg & Kuse Mental Rotations Test, RSPM = Raven's Standard Progressive Matrices; *M* = mean score or ability estimate, *SD* = standard deviation of sores or ability estimates, miss = rate of missing values, LR = p value of Andersen's likelihood ratio test (mean split), AUC = area under Receiver-Operating-Characteristic (ROC) curve, BIC = Bayes Information Criterion, Alpha = internal consistency.

latent factors while allowing for free estimation of all factor loadings.

Model fit in the hypothesized oblique three-factor-model with a higher-level factor was not acceptable with a comparative fit index (CFI) of 0.82 and a root mean square error of approximation (RMSEA) of 0.060. It is likely that model fit was negatively affected by MR items, showing relatively low standardized factor loadings, one of which (mr7) was non-significant ($p = .16$). Since the MR scale was also problematic in terms of unidimensionality according to Drasgow & Lissak's (1983) modified parallel analysis ($p = .01$), it is plausible that it's items do not load well on a single factor in the overall model. We therefore performed another CFA with just PC and SC items as indicators. The model clearly supported a two-factor solution with a higher-level factor, exhibiting good model fit (CFI = 0.97, RMSEA = 0.031). The model fit the data significantly better than a one-factor solution ($\chi^2(1) = 70.8$, $p < .001$) and a model with orthogonal factors and not higher-level factor ($\chi^2(1) = 22.0$, $p < .001$). All indicators showed significant positive factor loadings, with standardized coefficients ranging from 0.33 to 0.71. The results confirm that the items of the PC and SC scales load on two distinct but related factors. For graphical representations of the model structures including standardized factor loadings, variances and covariances, please see SI 5 and SI 6. Model fit statistics of all tested models can be found in Table 3.

3.2.2. Convergent validity

We used a structural equation model (SEM) to test for the correlation of two latent variables representing an overall *Minecraft*-based measure of intelligence and an overall paper-based measure of intelligence. As multivariate normality was not satisfied according to the Henze-Zirkler test ($HZ = 1.05$, $p = .008$), we used the MLM estimation procedure which is robust against violations of the normality assumption (Satorra & Bentler, 1994). The SEM contained a latent variable based on scale-level scores of the PC and SC scales on the one hand and another latent variable based on RSPM and VKMR scores on the other hand. MR was left aside, as it did not fit in the assumed factor structure. We expected a positive relationship between the latent variables. In accordance with our hypotheses, the SEM exhibited good model fit (robust CFI = .99) and the latent factors of *Minecraft*-based and paper-based tests showed a strong positive relationship ($r = 0.72$, $p = .002$). The model structure is depicted in Fig. 2. For measurement models of the PC and SC scales, please see SI 7.

Table 3
Factorial validity - CFA fit statistics.

Model	χ^2	df	CFI	90% CI RMSEA	SRMR	AIC
1	272	186	.82	[.044, .075]	.089	3083
2	390	189	.58	[.078, .10]	.10	3194
3	298	189	.77	[.052, .081]	.12	3103
4	84	75	.97	[.000, .061]	.058	1874
5	155	77	.76	[.068, .11]	.089	1941
6	106	77	.91	[.025, .078]	.12	1892

Note. Models: 1 = Three factors with higher level factor, 2 = One factor, 3 = Three orthogonal factors, 4 = Two factors with higher level factor (MR excluded), 5 = One factor (MR excluded), 6 = Two orthogonal factors (MR excluded),; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root mean square residual, AIC = Akaike's information criterion.

To test for convergent validity at subtest level we used Spearman's correlations, because some of the test scores were non-normal. The correlation between PC and RSPM was significant ($rs(118) = 0.28$, $p < .001$), but the effect size was somewhat lower than expected. The correlation between MR and VKMR was significant with a moderate effect size ($rs(118) = 0.39$, $p < .001$). The SC scale showed a significant correlation with VKMR ($rs(118) = 0.39$, $p < .001$). All other correlations ranged from $rs(118) = 0.20$ to $rs(118) = 0.46$ showing significant relationships between all pairs. A comprehensive overview of effect sizes can be found in Table 4. Subtest correlations based on simple sum scores without item elimination can be found in SI 8. Subtest correlations controlling for the order of test administration are reported in SI 9. The results were highly consistent with the results reported above.

3.2.3. Discriminant validity

To assess discriminant validity we used Pearson's and Filon's z-test (Diedenhofen & Musch, 2015) to test for differences between correlations expected to be high (e.g. MR and VKMR) and correlations expected to be comparatively low (e.g. MR and RSPM). The highest correlations were expected between PC and RSPM, both of which are matrix-based tasks (Gf), and between MR, SC and VKMR, all of which are spatial reasoning tasks (Gv). The correlations between PC and VKMR, MR and RSPM as well as SC and RSPM on the other hand were expected to be positive but significantly lower, as the tasks belong to related but distinct factors in the Cattell-Horn-Carroll model (Carroll, 1993; McGrew, 2005). Contrary to our hypothesis, the correlation between PC and VKMR was not significantly lower than the correlation between PC and RSPM ($z = -1.26$, $p = .90$). The correlation between MR and RSPM was lower than the correlation between MR and VKMR, but the difference was only marginally significant ($z = 1.33$, $p = .09$). In line with our hypothesis, the *Minecraft*-based SC scale showed its highest correlation with VKMR and its correlation with RSPM was significantly lower ($z = 2.08$, $p = .018$).

3.3. Effects of gender and minecraft experience

We used pairwise Welch tests and Wilcoxon rank-sum tests to determine the effect of gender and *Minecraft* experience on ability estimates at subscale-level. We did not expect gender or *Minecraft* experience to have an effect on performance in the PC task. However, we suspected a gender effect in the MR task and the SC task and an effect of *Minecraft* experience in the SC task. As expected, Welch's t-tests showed no significant group differences in the PC task (all $p > .10$). Furthermore, there were no significant group differences in the MR task (all $p > .14$). In the SC task, on the other hand, there were significant group differences with regard to gender ($t(104.9) = -3.361$, $p < .001$, $d = 0.62$) and prior *Minecraft* experience ($t(15.97) = -2.96$, $p = .005$, $d = 0.91$). For comparison, the conventional assessments did not exhibit any

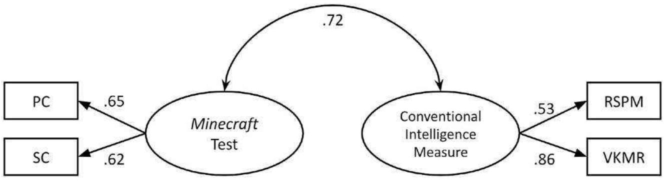


Fig. 2. Structural equation model (SEM) of the latent relationship between *Minecraft*-based and paper-based intelligence measures; PC = pattern completion task, SC = spatial construction task, VKMR = Vandenberg & Kuse Mental Rotations Test, RSPM = Raven's Standard Progressive Matrices.

Table 4
Convergent validity - Spearman's correlation coefficients.

	PC	MR	SC	VKMR	RSPM
PC	(.73)				
MR	.20*	(.62)			
SC	.41***	.20*	(.76)		
VKMR	.39***	.39***	.39***	(.90)	
RSPM	.28**	.28**	.20*	.46***	(.76)

Note. PC = pattern completion task, MR = mental rotation task, SC = spatial construction task, VKMR = Vandenberg & Kuse Mental Rotations Test, RSPM = Raven's Standard Progressive Matrices, *p < .05, **p < .01, ***p < .001; diagonal shows internal consistencies.

significant group differences (all p > .05).

No group differences were found with regard to enjoyment of the *Minecraft* testing experience between male and female participants (t (93.9) = 0.57, p = .57) and between participants with and without *Minecraft* experience (t(16.24) = −1.8384, p = .084). Furthermore, there were no group differences with regard to participants' positive attitude towards video games (all p > .075), but with regard to self-assessed gaming skills (all p < .020), as males and participants with *Minecraft* experience tended to rate their skills more favorably. Self-assessed gaming skills, in turn, were correlated with PC scores (rs(118) = 0.21, p = .024) and SC scores (rs(118) = 0.37, p < .001). Group differences with regard to gender in the SC task persisted when the effect of self-assessed gaming skills was partialled out (t(105) = −2.38, p = .019), but group differences with regard to *Minecraft* experience did not (t (15.6) = −1.98, p = .066). Similarly, the differences remained significant when correcting for time spent on the tutorial test, which was negatively correlated with SC task performance (rs(118) = −0.32, p < .001) and can be interpreted as an inverse proxy of actual *Minecraft* skills (gender: t(95.1) = −3.36, p = .001; *Minecraft* experience: t(14.3) = −2.16, p = .049).

All comparisons were reexamined using a Wilcoxon rank-sum text as a non-parametric alternative to the t-test. The Wilcoxon tests showed several additional group differences, namely an effect of gender on VKMR-scores (p = .035), as well as an effect of gender and *Minecraft* experience on positive attitude towards video games (both p < .046). Structurally, however, the results are comparable as SC remains the only *Minecraft*-based task where significance persists after a Bonferroni-correction for multiple comparisons. For a detailed overview of group differences with effect sizes and significance levels, please see Table 5.

3.4. Exploring behavioral data from game-logs

In order to explore whether the log data collected through the *Project Malmo* API encodes information about test takers' ability levels, we employed supervised machine learning techniques to predict sub-scale scores in the *Minecraft* test and the pen-and-paper tests, based on aggregate features derived from the logged variables. Considering the scope of the present paper we focus on log-data from the SC task, as the SC task provides the most opportunities for test takers to interact with the game environment, compared to the PC and MR tasks. Specifically, we used random forest regression models (Breiman, 2001) in conjunction with a nested cross validation scheme to assess how well the test scores could be predicted from the log data. Random forest models were fitted with 500 regression trees and up to 26 randomly selected features per split. In order to identify the optimal number of features used for each split we searched a hyper-parameter space including all even numbers between two and 26 and picked the model configuration that performed best on a validation set. This model was then evaluated on a testing set consisting of hold-out data. The inner loop of the cross-validation scheme used 10-fold cross-validation for hyper-parameter tuning, while the outer loop used Monte-Carlo cross validation with an 85/15 split and 30 iterations to estimate generalized model performance on testing data, which had not been previously used for training. The analyses were performed with the Caret (Kuhn, 2008) machine learning package for R.

Table 5
Group comparisons - gender and minecraft-experience.

	gender			Minecraft experience		
	d	*p*	*p**	*d*	*p*	*p**
PC	0.31	.10	.054	0.20	.60	.74
MR	0.20†	.14†	.16†	0.069	.79	.84
SC	0.62†	<.001†	<.001†	0.91†	.004†	.001†
RSPM	−0.14	.46	.87	−0.44	.23	.11
VKMR	0.32†	.05†	.035†	−0.17	.51	.53
TEQ	−0.11	.57	.57	0.54	.084	.066
PAVG	0.31	.098	.012	0.68	.075	.046
SAGS	0.57	.003	¡.001	0.84	.020	.006

Note. Results of Welch's t-tests and Wilcoxon rank-sum tests, p* indicates p value of Wilcoxon rank-sum tests, †denotes one sided hypothesis, positive effect sizes indicate higher sores for male participants or participants with prior *Minecraft* experience, respectively; PC = pattern completion task, MR = mental rotation task, SC = spatial construction task, RSPM = Raven's Standard Progressive Matrices, VKMR = Vandenberg & Kuse Mental Rotations Test, TEQ = test enjoyment questionnaire score, PAVG = positive attitude towards video games, SAGS = self-assessed gaming skills.

The features that were used for modelling capture how test takers interact with the game environment. We extracted the range, mean and standard deviation of each of the following variables: X, Y and Z coordinates of the avatar's position in space; X, Y and Z coordinates of the point in space that the cursor was focused on; the avatar's distance to the point in space that the cursor was focused on; the pitch (the vertical orientation of the avatar, i.e. looking up or down); the yaw (the horizontal orientation of the avatar, i.e. looking left or right). Additionally, we extracted overall distance travelled, as well as the time spent facing the model structure, the time spent facing the target structure, and the ratio of the two. Taken together, these 28 features provide a relatively complete picture of a test taker's idiosyncratic interactions with the virtual environment. We purposefully decided not to use metrics like the number of correctly placed blocks or completion times, because such data would have revealed too much information about the solution process and would have rendered the prediction problem trivial.

We found that the SC log-data was highly predictive of performance in the SC task, where the average correlation between predicted and observed test scores was r = 0.67. The SC log-data was less predictive of performance in the PC task, where the average correlation between predicted and observed test scores was r = 0.25, and in the MR task, where the average correlation between predicted and observed test scores was r = 0.20. Finally, the SC log-data predicted performance in the VKMR test, where the correlation between predicted and observed scores was on average r = 0.21, but not in the SPM test, where the average correlation was only r = 0.04. An overview of the results including error metrics can be found in Table 6.

To further explore the idea that the log-data can provide information about differentially successful solution strategies, which may be indic-

Table 6
Evaluation of performance prediction from SC log-data.

	RMSE		MAE		*r*	
	M	*SE*	*M*	*SE*	*M*	*SE*
SC	.77	.03	.64	.02	.67	.02
PC	1.18	.03	.95	.03	.25	.03
MR	.96	.09	.81	.08	.20	.11
VKMR	1.21	.03	1.00	.03	.21	.03
SPM	1.33	.03	1.08	.02	.04	.04

Note. Means and standard errors of evaluation metrics over 30 Monte-Carlo cross-validation iterations; RMSE = root mean squared error, MAE = mean absolute error, r = Pearson's correlation coefficient of predicted and observed test scores.

ative of different ability levels, we adopted an unsupervised learning approach. First, we used the Uniform Manifold Approximation and Projection (UMAP; McInnes, Healy, & Melville, 2018) algorithm in conjunction with density-based spatial clustering (DBSCAN; Ester, Kriegel, Sander, & Xu, 1996) to find clusters in the previously described 28-dimensional feature space. The clustering algorithm assigns data points to groups, such that similar data points are grouped together in the same cluster. The analysis revealed five distinct clusters (Fig. 3). We then analyzed between-cluster performance differences using analysis of variance (ANOVA) and a series of pairwise group comparisons. The ANOVA revealed significant differences in overall SC scores between clusters ($F(4,115) = 8.77$, $p < .001$). Cluster differences accounted for 23% of the variance in SC scores ($R^2 = 0.23$). Pairwise group comparisons revealed that cluster 3 in particular stood out for being associated with very low SC scores ($M = -3.20$, $SD = 1.93$), whereas cluster 1 was associated with particularly high scores ($M = 1.22$, $SD = 1.92$). Performance differences between cluster 2 ($M = -0.68$, $SD = 2.63$), cluster 4 ($M = 0.61$, $SD = 2.50$) and cluster 5 ($M = 0.85$, $SD = 2.41$) were less extreme. The results of all pairwise t-tests can be found in SI 10.

In order to gain a detailed understanding of the associations between specific behavioral patterns and SC test scores, we extracted feature importance scores (permutation importance; Breiman, 2001) from each of the random forest models evaluated in the outer cross-validation loop. The importance scores of each feature were averaged across all 30 cross-validation iterations. The results show that the features varied greatly in their predictive power. The avatar's position on the vertical axis (Y_Pos_Mean, Y_Pos_SD), the distance that was travelled within the game environment (Distance_Travelled), the mean distance to the point in the environment that the cursor was aimed at (Ray_Dist_Mean), and the range of the degree to which the avatar was oriented along the vertical axis (Pitch_Range) were identified as the most important predictors. A full ranking of the features by permutation importance can be found in SI 11. Importantly, the feature importance scores include complex non-linear effects and interactions, which prevents a directional interpretation of the scores. To investigate the non-linear relationships between the five most important features and SC test scores we fitted a series of local regression models using locally estimated scatter plot smoothing (LOESS; Cleveland, 1979). The resulting regression curves are depicted in Fig. 4. All of the depicted relationships are non-linear: while there is a strong positive relationship between Y_Pos_Mean and test scores for lower feature values, the relationship levels off for higher values; Y_Pos_SD shows a similar pattern; Ray_Dist_Mean, on the other hand, only shows a negative relationship for higher feature values; finally, Distance_Travelled and Pitch_Range show inverted-U shaped relationships indicating that mid-range values were associated with the highest test scores. The results suggest that specific behavioral profiles were associated with high scores in the SC task. Such behavioral profiles are characterized by efficient movement, appropriate distance to the focal objects in the environment, and high variability along the vertical spatial dimension. This means that successful participants were able to trade off spatial exploration and goal-directed behavior and effectively utilize all three spatial dimensions to solve the task. Importantly, when the same analysis was applied to the relationships between SC log-data and VKMR scores the associations were very similar to the ones reported above (see SI 12). This indicates that the identified features indeed capture individual differences in spatial reasoning and not just differences in participants' familiarity with the game controls or other artifacts. LOESS regression curves including confidence intervals for all features are reported in SI 13.

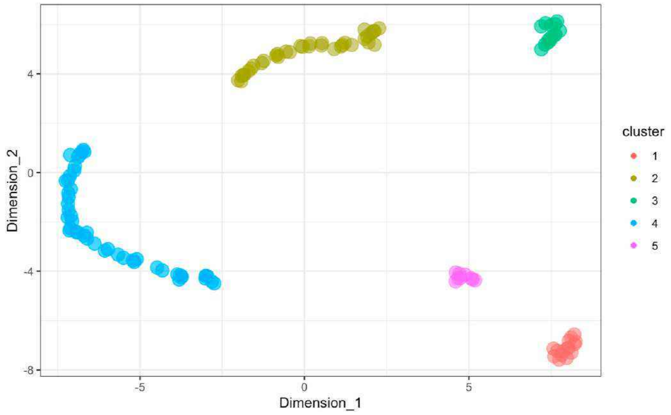


Fig. 3. Clusters in SC log-data projected to two-dimensional UMAP-space. Cluster assignments are represented by colors.

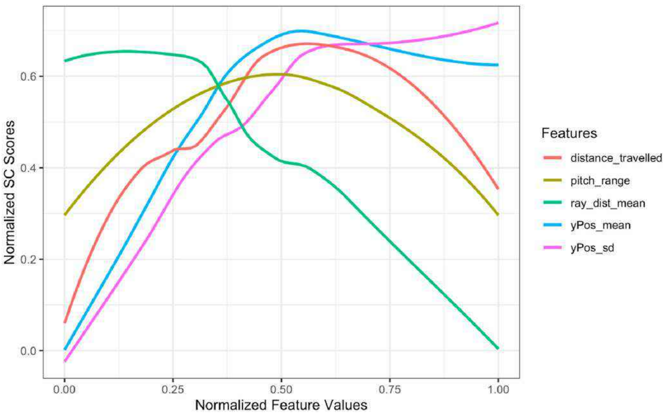


Fig. 4. LOESS regression curves of the relationships between the most important features used in the random forest models (by permutation importance) and SC test scores.

4. Discussion

4.1. Evaluation of the results

The aim of the present study was to investigate *Minecraft* and *Project Malmo* as a tool for game-based assessment research. We therefore set out to create measures of fluid intelligence and spatial ability in the virtual environment of the game. All three *Minecraft*-based subscales were successfully fitted to the Rasch model. Mental Rotation (MR) items, however, appeared to load on more than one factor in violation of the unidimensionality assumption. Internal consistencies were acceptable considering the low number of items per scale and the exploratory character of the project. When corrected for scale length (de Vet, Mokkink, Mosmuller, & Terwee, 2017), the reliabilities were comparable to those of other innovative game-based assessments proposed by Foroughi et al. (2016) and Quiroga et al. (2016). The results show that scale construction was least successful in the case of MR, the only *Minecraft*-based task with a multiple choice answer format, as reliability was comparatively low, there was little variance in item difficulties, and the scale was not unidimensional. The relatively low accuracy in combination with surprisingly low response times ($M = 18.23$, $Mdn = 13.52$, $SD = 17.23$) in the MR task resembles previous findings concerning effects of game-like task presentation. Washburn (2003) reports that participants responded faster and with lower accuracy when an otherwise identical cognitive task was framed as a game. The author concludes that the observed speed-accuracy trade-off was a characteristic effect of competition, i.e. participants trying to win the game. This interpretation is supported by the fact, that test takers on average spent less time on items that were answered incorrectly (see SI 14).

The problematic properties of the MR scale were reflected in the results of the CFA, revealing that the data did not support an overall three-factor solution with a higher level g-factor. This was due to the MR items not fitting in the hypothesized factor structure, a finding that is in line with the scale's violation of the unidimensionality assumption. A CFA with just Pattern Completion (PC) and Spatial Construction (SC) yielded good model fit for a two-factor solution with a higher level g-factor, suggesting that the scales measure distinct constructs. This result is in line with the literature supporting distinct – albeit related – factors for fluid reasoning and spatial abilities (Gf and Gv in the Cattell-Horn-Carroll model; McGrew, 2005; Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007). With regard to convergent validity, the *Minecraft*-based tests and the paper-based tests showed a strong correlation at the latent level between factors that can be described as nonverbal proxies for general intelligence made up of Gf and Gv. The latent relationship was lower than the one found by Quiroga et al. (2015) with a model including a wider range of game-based tasks from *Big Brain Academy* and a wider range of conventional measures of intelligence. The relationship was in the same range as the correlations found by Foroughi et al. (2016), which were based on a single video game, *Portal 2*, and measures of fluid intelligence only. At subscale level, the *Minecraft*-based tests correlated moderately with each other and their paper-based counterparts. The correlations were comparable in size to correlations between different WISC and WAIS subtests for average to high IQ participants (Detterman & Daniel, 1989) and the subtest-level correlations between game-based assessments and conventional tests reported by Quiroga et al. (2016) or Jones et al. (1986). However, they were lower than what would be expected for conventional tests aiming to measure similar constructs and also somewhat lower than the subtest-level correlations reported by Quiroga et al. (2015). The results concerning discriminant validity were mixed. As expected, the SC scale showed a significantly lower correlation with RSPM than with VKMR. The MR scale also showed the expected pattern, but the difference of the correlation coefficients was only marginally significant. The expected pattern of correlations was not found in the case of the PC scale, which did not show a lower correlation with VKMR than with RSPM. The last finding is not in line with our hypotheses, but on the other hand previous research finds that matrix tasks like Raven's Advanced Progressive Matrices involve spatial ability and show moderate correlations with mental rotation tests (Mackintosh & Bennett, 2005; Schweizer et al., 2007). It is possible that the relationship with spatial ability is even more accentuated in the PC task, since the stimuli are three-dimensional and the item transformations are based on movement patterns of blocks and changes of location rather than transformations of geometric shapes in a single matrix as in RSPM. This interpretation is supported by findings that scores in a cognitive task showed an increased association with Gv when presented in a game-like fashion (McPherson & Burns, 2007).

Group differences with regard to gender and *Minecraft* experience show a male advantage and a positive effect of *Minecraft* experience on ability estimates in the SC task, even when corrected for self-assessed gaming skills or actual *Minecraft* skills demonstrated in the tutorial test. The findings are in line with our hypotheses and prior research suggesting gender differences in spatial reasoning (Voyer et al., 1995). The fact that prior findings regarding gender differences in spatial reasoning were reproduced can be interpreted as additional evidence of construct validity (Cronbach & Meehl, 1955). On the other hand, such differences raise questions regarding the fairness of the test. Given the small number of students without *Minecraft* experience (only 14 individuals), the effect of prior *Minecraft* experience has to be treated with caution. Assuming, however, that the effect is reliable, there are at least three possible explanations. First, it is possible that participants' lack of familiarity with the game controls interferes with the solution process – either by preventing them from submitting the correct solution through psychomotor constraints or by absorbing cognitive resources that could otherwise be used to solve the problem (see Sweller, 2010). This interpretation would pose a threat to the fairness of the test, as individuals without *Minecraft* experience would be disadvantaged. Second, exposure to video games, in this case *Minecraft*, may be the cause of a real improvement in spatial reasoning among the subsample with *Minecraft* experience (Granic, Lobel, & Engels, 2014; Uttal et al., 2013). Finally, there is the possibility that the results are in part attributable to reverse causality, as individuals with good spatial ability may feel drawn to games where they can put their spatial strength to use and thus may be more likely to have experience with *Minecraft*. While a single causal explanation of the effects is highly unlikely it should be noted that the latter two explanations are not supported by our data as there was no positive relationship between *Minecraft* experience and VKMR scores. Hence, it is likely that familiarity with *Minecraft* gives test takers an advantage in the SC task.

The analysis of game log-data indicates that testing performance can be predicted from relatively abstract patterns in test takers' interactions with the game environment. Specifically, we found that SC scores can be predicted accurately on hold-out data, and that different behavioral clusters were associated with performance differences in the SC task. These results are in line with earlier findings indicating that performance in the wells-task can be predicted from discrete action sequences (Hao et al., 2015; Zhu et al., 2016) and that cluster analyses can pick up on performance related information (Bergner et al., 2014). Importantly, in distinction from previous research, we also tested how well the log-data predicted performance in the other *Minecraft* tasks and the pen-and-paper tasks in order to arrive at a first assessment of convergent and discriminant validity. The results showed that the SC log data was to a lesser extent predictive of performance in the PC and MR tasks compared to SC performance, indicating that it encodes task-specific information and not just navigation skills or familiarity with the game controls. Interestingly, the SC log-data was also more predictive of VKMR scores compared to RSPM scores (i.e., the quality of the predictions mapped onto the strength of the associations that would be expected based on the CHC), which indicates that the log-data indeed captures information about participants' spatial reasoning ability. The observed pattern of results hence provides additional evidence for the construct validity of the *Minecraft* test. Another key difference that sets the *Minecraft* test apart from previous work is the volume and granularity of data that is collected. While previous research (e.g. Bergner et al., 2014; Hao et al., 2015; Shu et al., 2017; Zhu et al., 2016), mainly dealt with discrete univariate process data, such as action sequences, the *Minecraft*-based assessment opens up a larger, less well-defined action space and produces continuous multivariate data, potentially encoding more information, but also posing methodological challenges that have to be addressed in future research.

Our initial investigation of the relationships between individual behavioral features and test scores revealed distinct behavioral profiles that are associated with high levels of spatial ability. Such behavioral profiles are characterized by efficient movement within the game environment, appropriate distance to the focal objects within the environment, and high variability along the vertical spatial dimension. The inverted U-shaped relationships between movement, as well as distance to focal objects, and test scores speak to participants' ability to trade off spatial exploration and goal-directed behavior. While travelling further distances and creating distance between the avatar and focal objects can help to explore relevant parts of the environment, such behaviors may interfere with the construction of the target structure, which requires test takers to focus on a particular part of the environment and to be close enough to place blocks in the designated area. These findings highlight the known relationship between intelligence and goal-directed behavior (Duncan, Emslie, Williams, Johnson, & Freer, 1996) in the context of spatial reasoning and relate loosely to the process coordination view of spatial reasoning (Pellegrino, Alderton, & Shute, 1984). The positive relationships between vertical movement and test scores indicate that individuals who changed their elevation levels were particularly successful. Associated behaviors include jumping up and down, building scaffolds in order to explore the model structure and attach blocks to the target structure, or to construct the target structure while standing on top of it. Such behaviors speak to participants' ability to utilize all three spatial dimensions when solving the task. This observation is consistent with previous research linking flexible strategy choice to performance in spatial reasoning (Hegarty, 2010). Furthermore, our results are in line with previous findings indicating that high levels of spatial ability result in the adoption of more holistic solution strategies leading to superior performance in spatial reasoning tasks (Buckley, Seery, & Canty, 2019). Our results also can also be interpreted in light of Hegarty's (2010) idea of representational meta-competence as a component of spatial reasoning. Meta-representational competence (diSessa, 2004) describes peoples' ability to choose the best representation for a particular task and interact effectively with novel displays without explicit instruction. While the behavioral log-data does not enable us to directly assess test-takers' representational states, the data clearly suggests differences in the effectiveness of test-takers' approach to the virtual environment and the task, which are consistent with the concept of meta-representational competence. Importantly, the relationships between SC log-data and VKMR scores closely mirrored the relationships between SC log-data and SC scores, suggesting that the identified features indeed capture individual differences in spatial reasoning and not just differences in participants' familiarity with the game controls or other artifacts. Taken together, these findings provide further support for the validity of the SC task and demonstrate how the analysis of multivariate process data can contribute to our understanding of the psychometric properties of game-based assessments. As such, our findings provide a meaningful addition to earlier work on the analysis of process data in complex problem solving tasks (Bergner et al., 2014; Hao et al., 2015; Shu et al., 2017; Zhu et al., 2016) and recent research exploring the use of sensing technologies for the collection of process data in block design tasks (Cha, Ainooson, & Kunda, 2018; Lee et al., 2018).

With regard to task design and presentation in game-based assessments, the present study raises four major points. First, multiple choice formats as in MR seem to encourage guessing in game-like tasks, as test takers may feel encouraged to explore and use trial and error strategies – possibly because it is less clear to students that they are taking a test, compared to conventional tests on paper (Washburn, 2003). Second, complex constructed response formats as in the SC task eliminate the danger of excessive guessing, but may introduce other distortions as a result of differences in psychomotor abilities (Foroughi et al., 2016) and familiarity with computer games (Hambrick, Oswald, Darowski, Rench, & Brou, 2010). We agree with prior research (Foroughi et al., 2016; Greiff et al., 2016) that test performance needs to be independent of gaming experience and computer skills. We therefore want to stress the importance of tasks with low psychomotor requirements and tutorials specifically preparing test takers for the tasks they are about to face. Tutorials are essential, not only to offset differences in familiarity with game controls, but also to balance out differential learning curves in video games (Jones, 1984). Another potential solution would be the presentation of tasks in virtual reality (VR), which is possible in *Minecraft* as it has already been adapted to be playable on consumer VR devices like the Oculus Rift. Participants without gaming experience may find it more intuitive to look around by moving their head in VR than they would looking around with the mouse. Third, the effects of three-dimensional, game-like presentation on performance, motivation and the factorial structure of intelligence measures should be further investigated, as there is the possibility that game-like presentation differentially affects test takers' motivation (Mislevy et al., 2014) and, as illustrated by the PC task, may disproportionally emphasize the importance of Gv (McPherson & Burns, 2007). Finally, our results indicate that complex tasks, such as the SC task, yield useful process data that captures information about test takers ability and could be used alongside with traditional scoring approaches in future assessments.

4.2. Limitations and directions for future work

The present study has some limitations with regard to sample size and representativity of participants and psychometric tasks. In future studies, the assessment should be recalibrated and cross-validated with a wider range of tasks, a larger sample of participants, and a larger number of items per scale, allowing for the creation of item banks for adaptive testing and an in depth investigation of differential item functioning. Additional conventional measures of intelligence could be used for the purpose of construct validation, for example measures of Gc could be included for a better assessment of discriminant validity. In the context of construct validation it may also be advisable to conduct a multi-trait-multi-methods analysis. Most importantly, criterion validity should be addressed by including external outcomes such as academic achievement, specifically in the fields of science, technology, engineering and mathematics (STEM), which are closely associated with fluid intelligence and spatial reasoning (Kell & Lubinski, 2013; Wai, Lubinski, & Benbow, 2009). Another key point is incremental validity, the question of whether the game-based tests predict criterion variance not shared with paper-based tests, as it would indicate the involvement of ability factors that are not present in conventional assessments (Jones et al., 1986). We found anecdotal evidence that a student with autism particularly enjoyed the game-based assessment whereas the same student found it difficult to complete paper-based assessments. This indicates that games have the potential to bring out capabilities in certain students that they do not demonstrate on paper. This is worthy of follow-up study. Future studies should also take into account the role of the narrative and other game design principles. For example, it would be possible to give test-takers more control about the testing experience by enabling non-linear game-play sequences, where players determine the order in which they work through the tasks (Kim & Shute, 2015).

Given the overall favorable results of our study, it seems appropriate to briefly discuss two opportunities for future work, which set the current project apart from previous work on game-based assessment: the analysis of log-data from the game environment and automatic item generation. As our exploratory analyses suggest, the log-data captures information about test-takers' performance. Future work could expand more on this finding. First, since the present paper focuses on SC log-data, future work should explore the properties of the log-data from the remaining tasks. Second, log-data could be used to detect differential strategies and problematic behaviors such as rapid guessing and refusal to work, or problems with the game controls. Third, it would be interesting to predict not only ability estimates, but also external criteria from the log-data. This could provide additional evidence that the data encodes information about test-takers' ability levels. As mentioned above, the complexity of the log-data also calls for methodological innovation, as the structure of the data is not compatible with conventional psychometric modelling approaches. Furthermore, the *Minecraft* environment seems predestined for automatic item generation - employing computer algorithms to generate new items on-the-fly (Gierl & Haladyna, 2013; Gierl and Lai, 2012). Due to the block-based organization of the environment, structures in *Minecraft* can easily be represented as arrays in various programming languages. These data structures can be transformed according to computer-generated rules and translated back into code that is processed by the *Project Malmo* API. All of this can happen in a single Python script. That means that an automatic item generator can produce fully functioning items that are automatically implemented in the *Minecraft* environment and can be run in real time.

4.3. Conclusions

All in all, the present study suggests that customized assessments based on existing games have the potential to become a viable companion to conventional assessments in the future. We have shown for the first time that *Minecraft* can be used as an assessment platform and that reasoning tests can be implemented in the three-dimensional game environment. We found that young children are not necessarily distracted by the game-like features of *Minecraft* and genuinely engage with the assessment task. Unlike previous work on the measurement of intelligence with video games, the present study uses a Rasch model for scale construction, which is a first step towards the application of computer adaptive testing in game-based assessment. Adaptivity is especially important in game-based assessment, as it allows for the presentation of moderately difficult items relative to individuals' ability levels and moderate task difficulty is one of the key principles of game design (Gee, 2003; Przybylski, Rigby, & Ryan, 2010). Hence, adaptivity is likely to increase enjoyment and motivation in addition to test efficiency (Mislevy et al., 2014). The flexibility of *Project Malmo*, enabling the definition of customized task types and performance criteria, combined with its capacity to record real-time process data and its suitability for automatic item generation, makes *Minecraft* a very promising tool for future work in game-based assessment. Furthermore, the implementation of intelligence tasks in an environment that is open to machine learning applications - and thus also computational models of cognition - has the long term potential to integrate the research areas of psychometric task design and cognitive theory, which is a key challenge in the field of game-based assessment and psychometrics in general (Embretson, 1998; Primi, 2014; Quiroga et al., 2016).

Acknowledgements

We thank Frank Spohn, Jan Wacker and Jana Mangels for their valuable comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2021.106701.

References

Arce-Ferrer, A. J., & Martínez Guzmán, E. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the raven standard progressive matrices test. *Educational and Psychological Measurement, 69*(5), 855–867.

Baniqued, P. L., Lee, H., Voss, M. W., Basak, C., Cosman, J. D., DeSouza, S., et al. (2013). Selling points: What cognitive abilities are tapped by casual video games? *Acta Psychologica, 142*(1), 74–86.

Bergner, Y., Shu, Z., & von Davier, A. (2014). Visualization and confirmatory clustering of sequence data from a simulation-based assessment task. *Proceedings of the 7th International Conference on Educational Data Mining*, 177–184.

Borghans, L., Meijers, H., & ter Weel, B. (2013). The importance of intrinsic and extrinsic motivation for measuring IQ. *Economics of Education Review, 34*, 17–28.

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

Buckley, J., Seery, N., & Canty, D. (2019). Investigating the use of spatial reasoning strategies in geometric problem solving. *International Journal of Technology and Design Education, 29*(2), 341–362.

Buford, C. C., & O'Leary, B. J. (2015). Assessment of fluid intelligence utilizing a computer simulated game. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 1–17.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Cha, S., Ainooson, J., & Kunda, M. (2018). *Quantifying human behavior on the block design test through automated multi-level analysis of overhead video*. arXiv: 1811.07488.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829–836.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 9–15.

Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence, 13*(4), 349–359.

diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction, 22*(3), 293–331.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One, 10*(4), 1–12.

Dikkers, S. (2015). *Teachercraft: How teachers learn to use Minecraft in their classrooms*. Pittsburgh, Pa: ETC Press.

Dorner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8*, 1–11.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363–373.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*(19), 7716–7720.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology, 30*(3), 257–303.

Ekaputra, G., Lim, C., & Eng, K. I. (2013). Minecraft: A game as an education and scientific learning tool. In *Information systems international conference (ISICO)* (pp. 237–242).

Ellison, T. L., Evans, J. N., & Pike, J. (2016). Minecraft, teachers, parents, and learning: What they need to know and understand. *School Community Journal, 26*(2), 25–43.

Embretson, S. (1998). A cognitive design system Approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences* (1st ed. edition). Routledge.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on Knowledge discovery and data mining, KDD'96* (pp. 226–231). Portland, Oregon: AAAI Press.

Flanagan, D. P., & Dixon, S. G. (2014). The cattell-horn-carroll theory of cognitive abilities. In C. R. Reynolds, K. J. Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education*. American Cancer Society.

Foroughi, C. K., Serraino, C., Parasuraman, R., & Boehm-Davis, D. A. (2016). Can we create a measure of fluid intelligence using Puzzle Creator within Portal 2? *Intelligence, 56*, 58–64.

Fu, J., Zapata, D., & Mavronikolas, E. (2014). Statistical methods for assessments in simulations and serious games: Statistical methods in simulations and serious games. *ETS Research Report Series, 2014*(2), 1–17.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy* (1 edition edition). New York: Palgrave Macmillan.

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation: Theory and practice*. New York: Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273–298.

Godwin, K. E., Lomas, D., Koedinger, K. R., & Fisher, A. V. (2015). Monster mischief: Designing a video game to assess selective sustained attention. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 18–39.

Golino, H. F., & Gomes, C. M. A. (2016). Random forest as an imputation method for education and psychology research: Its impact on item fit and difficulty of the Rasch model. *International Journal of Research and Method in Education, 39*(4), 401–421.

Granic, I., Lobel, A., & Engels, R. C. M. E. (2014). The benefits of playing video games. *American Psychologist, 69*(1), 66–78.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: The MicroDYN approach. In F. Scheuermann, & J. Björnsson (Eds.), *The Transition to Computer-Based Assessment. Lessons learned from large-scale surveys and implications for testing* (pp. 157–163). Luxembourg: Office for Official Publications of the European Communities.

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology, 106*(3), 666–680.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36–46.

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contextsâ€"Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*(2), 364–379.

Haier, R. J., Siegel, B., Tang, C., Abel, L., & Buchsbaum, M. S. (1992). Intelligence and changes in regional cerebral glucose metabolic rate following learning. *Intelligence, 16*(3), 415–426.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology, 24*(8), 1149–1167.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7*(1), 18.

Hegarty, M. (2010). Components of spatial intelligence. In *Psychology of learning and motivation, volume 52 of the psychology of learning and motivation* (pp. 265–297). Academic Press.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75*(3), 242.

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 41–68). New York, NY, US: The Guilford Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253–270.

Hunt, E. (2011). *Human intelligence*. New York: Cambridge University Press.

Hunt, E., & Pellegrino, J. (1985). Using interactive computing to expand intelligence testing: A critique and prospectus. *Intelligence, 9*(3), 207–236.

Jackson, D. N., Vernon, P. A., & Jackson, D. N. (1993). Dynamic spatial performance and general intelligence. *Intelligence, 17*(4), 451–460.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*(1), 1–15.

Johnson, M., Hofmann, K., Hutton, T., & Bignell, D. (2016). The Malmo platform for artificial intelligence experimentation. In *IJCAI* (pp. 4246–4247).

Jones, M. B. (1984). Video games as psychological tests. *Simulation & Games, 15*(2), 131–157.

Jones, M. B., Dunlap, W. P., & Bilodeau, I. M. (1986). Comparison of video game and conventional test performance. *Simulation & Games, 17*(4), 435–446.

Karsenti, T., Bugmann, J., & Gros, P. (2017). *Transforming education with Minecraft? Results of an exploratory study conducted with 118 elementary-school students*. Montréal: CRIFPE.

Kell, H. J., & Lubinski, D. (2013). Spatial ability: A neglected talent in educational and occupational settings. *Roeper Review, 35*(4), 219–230.

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.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(1), 1–26.

Lee, K., Jeong, D., Schindler, R. C., Hlavaty, L. E., Gross, S. I., & Short, E. J. (2018). Interactive block games for assessing children's cognitive skills: Design and preliminary evaluation. *Frontiers in Pediatrics, 6*.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games, 4*(2).

Mackintosh, N., & Bennett, E. (2005). What do Raven's matrices measure? An analysis in terms of sex differences. *Intelligence, 33*(6), 663–674.

Mair, P., Hatzinger, R., & Maier, M. J. (2009). Extended Rasch modeling: The R package eRm. *Journal of Statistical Software, 20*.

Mair, P., Reise, S. P., & Bentler, P. M. (2008). *IRT goodness-of-fit using approaches from logistic regression*.

Mané, A., & Donchin, E. (1989). The space fortress game. *Acta Psychologica, 71*(1–3), 17–22.

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

McGrew, K. S. (2005). The cattell-horn-carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136–181). New York, NY, US: Guilford Press.

McInnes, L., Healy, J., & Melville, J. (2018). *Umap: Uniform Manifold approximation and projection for dimension reduction*. arXiv: 1802.03426.

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

McPherson, J., & Burns, N. R. (2008). Assessing the validity of computer-game-like tests of processing speed and working memory. *Behavior Research Methods, 40*(4), 969–981.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Meijer, J., & Oostdam, R. (2007). Test anxiety and intelligence testing: A closer examination of the stage-fright hypothesis and the influence of stressful instruction. *Anxiety, Stress & Coping, 20*(1), 77–91.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., et al. (2014). *Psychometric considerations in game-based assessment*. Charleston: CreateSpace Independent Publishing Platform.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

National assessment of educational progress - technology and engineering literacy assessment. (2014). https://nces.ed.gov/nationsreportcard/tel/.

Oostdam, R., & Meijer, J. (2003). Influence of test anxiety on measurement of intelligence. *Psychological Reports, 92*(1), 3–20.

Pellegrino, J. W., Alderton, D. L., & Shute, V. J. (1984). Understanding spatial ability. *Educational Psychologist, 19*(4), 239–253.

Peters, M., Lang, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance. *Brain and Cognition*, (28), 39–58.

Primi, R. (2014). Developing a fluid intelligence scale through a combination of Rasch modeling and cognitive psychology. *Psychological Assessment, 26*(3), 774–788.

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology, 14*(2), 154–166.

Pusey, M., & Pusey, G. (2015). Using Minecraft in the science classroom. *International Journal of Innovative Science and Modern Engineering, 23*(3), 22–34.

Quiroga, M., Diaz, A., Román, F., Privado, J., & Colom, R. (2019). Intelligence and video games: Beyond "brain-games". *Intelligence, 75*, 85–94.

Quiroga, M., Escorial, S., RomÃ¡n, F. J., Morillo, D., Jarabo, A., Privado, J., et al. (2015). Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can! *Intelligence, 53*, 1–7.

Quiroga, M., 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.

Quiroga, M., Román, F., Catalán, A., Rodríguez, H., Ruiz, J., & Herranz, M. (2011). Videogame performance (not always) requires intelligence. *International Journal of Online Pedagogy and Course Design, 1*(3), 18–32.

Quiroga, M., Román, F. J., De La Fuente, J., Privado, J., & Colom, R. (2016). The measurement of intelligence in the XXI century using video games. *Spanish Journal of Psychology, 19*, 1–13.

Rabbitt, P., Banerji, N., & Szymanski, A. (1989). Space fortress as an IQ test? Predictions of learning and of practised performance in a complex interactive video-game. *Acta Psychologica, 71*(1–3), 243–257.

Raven, J., Raven, J., & Court, J. (2000). *Manual for Raven's progressive matrices and vocabulary scales. Section 3, Standard progressive matrices (including the Parallel and Plus versions)*. Oxford: OPP Ltd.

Rosseel, Y. (2012). Lavaan : An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

van Rossum, G. (1995). *Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI)*. Amsterdam, May 1995.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA, US: Sage Publications, Inc.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan, & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). New York, NY, US: The Guilford Press.

Schweizer, K., Goldhammer, F., Rauch, W., & Moosbrugger, H. (2007). On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Personality and Individual Differences, 43*(8), 1998–2010.

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling*, (59), 109–131.

Smith, E. V., & Smith, R. M. (Eds.). (2004). *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove, Minn: JAM Press.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123–138.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., et al. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352–402.

Ventura, M., Shute, V., Wright, T., & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology, 4*.

de Vet, H. C., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearmanâ"Brown prophecy formula and cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology, 85*, 45–49.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250–270.

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, & Computers, 35*(2), 185–193.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817–835.

Zhu, M., Shu, Z., & von Davier, A. (2016). Using networks to visualize and analyze process data for educational assessment: Network analysis for process data. *Journal of Educational Measurement, 53*(2), 190–211.