Acta Psychologica 259 (2025) 105252

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Evaluating planning through play: Exploring the use of mini games to assess planning abilities

Emma G. Cunningham [a], Daphné Bavelier [b,c,*], C. Shawn Green [a]

[a] Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, United States
[b] Faculty of Psychology and Educational Sciences, University of Geneva, Boulevard du Pont d'Arve 40, 1211 Geneva 4, Switzerland
[c] Fondation Campus Biotech, Geneva, Switzerland

ARTICLE INFO

Keywords:
Cognition
Planning
Video games
Stealth assessment
Look ahead

ABSTRACT

Planning, or the ability to generate, organize, and implement a sequence of steps toward a goal, is essential for success across a wide range of activities, from preparing a meal to developing a software program. Indeed, robust planning abilities have been found to predict math achievement in children and support independent living in older adults. However, common tasks used to measure planning often fail to correlate with one another, suggesting they may not assess the same underlying skill. To explore a novel approach to measuring planning, this study examined performance on four planning mini games, a non-planning control game, and a battery of cognitive tasks measuring related cognitive skills, including three standard planning tasks. As hypothesized, the planning mini games showed stronger intercorrelations than previously shown between traditional tasks, suggesting they may capture a more consistent and unified planning construct. Notably, two of the selected mini games emerged as particularly promising paradigms for assessing planning skills. These findings provide initial evidence that mini games such as those explored here could complement or replace traditional cognitive planning tasks, offering an appropriately complex evaluation of the multifaceted skill of planning.

1. Introduction

Video games are a ubiquitous form of entertainment. Billions of individuals worldwide play video games in some capacity, with video game play cutting across essentially all major demographic categories, including gender, race, and age (De Schutter, 2011; Stephen & Edwards, 2017). The global appeal of video games stems at least partially from their ability to motivate players by providing intrinsically valuable to-be-reached goals and achievements (Jones, 1984). Together such motivational systems inspire players to invest considerable amounts of time in gaming, in some cases into the hundreds of hours in a single game (Gee, 2007; Plass et al., 2020).

Critically, the way in which those in-game goals and achievements are obtained is through the development and utilization of various skills and abilities, from the ability to rapidly react and press a sequence of buttons, to the ability to mentally visualize certain spatial patterns, to the ability to work with others toward a common goal (Gee, 2003, 2007; Kenwright, 2023). The extent to which a video game necessitates that the player effectively utilizes a skill or ability is referred to by Bowman (2021) as a “demand.” More specifically, Bowman (2021) posits that video game play may involve one or more of four key types of demand: cognitive, emotional, physical, and social. Of particular relevance to the current work is the capacity of video games to place significant demands on a host of cognitive functions. Indeed, there is a great deal of evidence, derived from studies using many different types of methods (e.g., survey work, correlational analyses, etc.), indicating that video games have the capacity to load on a wide variety of cognitive functions depending on the particular mechanics inherent in the games (Bavelier & Green, 2019; Bediou et al., 2023; Bowman, 2021; Chisholm & Kingstone, 2015; Cunningham & Green, 2023; Dale et al., 2020; Feng et al., 2007; Green et al., 2012; Kim et al., 2017; Zhang et al., 2021). These cognitive functions include selective and sustained attention, executive functions, spatial cognition, reasoning and decision-making, multitasking and task-switching, and low-level perception, among others.

The capacity of video games to place clear demand on various cognitive functions is a feature that is shared with psychological tasks that are specifically designed to measure the given cognitive functions. Psychological tasks, however, often attempt to place demand on a single cognitive function while, at the same time, intentionally minimizing all other extraneous factors. This serves to isolate the cognitive function of

* Corresponding author at: Faculty of Psychology and Educational Sciences, University of Geneva, Boulevard du Pont d'Arve 40, 1211 Geneva 4, Switzerland.
E-mail address: daphne.bavelier@unige.ch (D. Bavelier).

https://doi.org/10.1016/j.actpsy.2025.105252
Received 26 March 2025; Received in revised form 13 June 2025; Accepted 2 July 2025
Available online 18 July 2025

interest from all other potentially confounding elements. In other words, most psychological tasks are intentionally sterile (Ma et al., 2022; van Opheusden & Ma, 2019).

It has long been argued that such sterile single-cognitive-function-targeting tasks are not a good model of how those cognitive functions work in real-world environments where there is, for instance, less direct repetition, more variety, a greater need to decide when and how to deploy each cognitive function, etc. (Newell, 1973; van Opheusden & Ma, 2019). In line with this perspective, our goal is not merely to isolate planning as a discrete construct, but to examine how it operates within dynamic, context-rich environments that more closely mirror the complexity of everyday situations. In this latter light, video games arguably provide a superior context for the measurement of cognitive functions than isolated and sterile tasks, as they require that individuals utilize those functions in environments that, while not entirely real-world, are nonetheless more complex and dynamic.

Consistent with this idea, there exists a wealth of research focused on using video games for cognitive skill assessment (Jones, 1984; Rosas et al., 2015; Shute et al., 2016; Shute & Rahimi, 2021). This style of assessment is sometimes referred to as *stealth assessment*, as it is administered seamlessly within an inconspicuous evaluation environment (Shute & Ventura, 2013). Games used to assess skills do so by harnessing embedded measurable actions within gameplay and recording outcome variables of interest natively within the game environment. For example, Shute et al. (2016) measured problem-solving skills via a modified version of the popular mobile game *Plants* vs. *Zombies 2* (Popcap Games and Electronic Arts). Various in-game actions (e.g., the use of a high-cost tool to stop enemies in either high or low danger situations) were categorized based on their alignment with several facets of problem-solving. Critically, those game-derived variables were then found to correlate with external measures of problem-solving, indicating the measures were valid.

Games used in place of traditional cognitive assessments for a variety of skills have been found to be as effective at measuring the skill of interest as more conventional means, while also being rated as more enjoyable and engaging to participants than traditional assessment tools (Shute et al., 2016; Shute & Wang, 2015; Tenorio Delgado et al., 2016; Ventura, Shute, Wright, & Zhao, 2013). For example, Baniqued et al. (2013) utilized a suite of games called mini games, which are smaller scale and simpler in terms of mechanics than traditional video games, but which share the basic features found in all video games. In this study, the authors first made a set of a priori predictions, based on the game characteristics, as to which cognitive abilities would be under the most demand in each of the games in the suite. They then compared participants' game-based performance to performance on traditional measures of those cognitive functions. As expected, they found that performance on those games that they had a priori categorized as placing demand on certain cognitive functions in turn correlated with performance on the standard psychological tests of those functions. Similarly, Ventura and Shute (2013) assessed participants level of intellectual persistence through both a stealth assessment game–a game-based measure of persistence embedded in a physics playground game–and a traditional measurement task (Eisenberger & Leonard, 1980; Ventura, Shute, & Zhao, 2013). They reported a strong correlation between these two measures, indicating that game-based persistence measurement and task-based outcomes align even when controlling for the effects of pre-test score, gender, and video game playing experience.

Importantly, the use of video games as stealth assessments may actually serve to increase the accuracy of the measurements, particularly in individuals who are unable to perform to their maximum potential in more sterile, boring, and/or intimidating task setups. To this point, Rosas et al. (2015) compared kindergarten through third grade participants' performance on standard cognitive tasks measuring intelligence, arithmetic, and reading skills with their performance on games assessing the same skills. They found that, although the overall scores between the two measurement methods correlated as expected, participants with lower academic performance scored higher on the game-based assessment methods compared to the traditional tests. Such a result could be attributed to a host of mechanisms including an increase in effort driven by intrinsically motivating game features and/or a reduction in stereotype threat (i.e., where more formal psychological measurements may activate certain negative stereotypes about certain groups, whereas video games are less likely to do so; Croizet & Claire, 1998; Nguyen & Ryan, 2008).

A domain whose history suggests video game-based stealth assessment may be particularly aided by the use of such an approach is the measurement of planning skills. Indeed, many of the most prominent tasks utilized in this literature are exceptionally sterile and arguably over-constrained. An example is the Tower of Hanoi, which requires rearranging discs on pegs to match a goal arrangement while following a number of rigid rules (e.g., move one disc at a time, don't place a larger disc on a smaller disc, at least one disc must be on each peg at all times, etc.). Overly sterile and constrained tasks of this type do not allow participants to navigate the problem space in a naturalistic manner, and further, by constraining the possible routes a participant can take to solve a problem, may obscure important details of the underlying planning process (van Opheusden & Ma, 2019).

The issues associated with the simplicity, rigidity, and sterility of many tasks in the planning literature have long been recognized by those in the field (Shallice & Burgess, 1991; van Opheusden & Ma, 2019). One route that researchers have taken to address this issue has been to create or adapt existing tasks with the aim of increasing the number of possible actions available to a participant in a given problem space. One can think of this approach as using many of the traditional approaches above as a foundation (i.e., remaining somewhat task-based and sterile), but pushing the envelope away from simple rule-based solutions and toward greater degrees of complexity. This approach aligns with the common objective of comparing machine learning algorithms to human planning, as many of the tasks that have been developed via this perspective present a challenging paradigm for humans and machine learning agents alike. For example, expanding the game Tic-Tac-Toe to require four, rather than three pieces in a row to win, significantly increases the task's complexity, while still remaining amenable to computational modeling of participant behavior (Lake et al., 2015; van Opheusden & Ma, 2019).

In a separate approach, rather than starting with the goal of computational tractability and moving toward greater degrees of complexity from that anchor point, other researchers have instead started by considering real-world situations where people need to plan and then working backward to convert those into more tractable tasks (Shallice & Burgess, 1991). For example, in the Hayes-Roth (1980) planning task, participants were given the following information: a list of errands to complete in a fictional town, a map of the town, starting and ending locations, and the total amount of time they had to complete their errands (which was less than the amount of time necessary to complete all of the errands). Tasks in this vein, while they clearly share some components of real-world planning, often suffer from another issue, which is that they rely too heavily on the specific context of the task (e.g., an arbitrary set of errands in a fictional shopping area with unusual constraints). Such tasks may primarily capture performance driven by idiosyncratic factors, especially when they force participants away from how they normally plan, or conversely, promote strategy usage which may not represent true planning capacity in novel scenarios (Burgess et al., 2005). And indeed, in the task above, most participants initially adopted the strategy of using an efficient route between errands (i.e., trying to reduce the distance traveled from errand to errand). However, given the particular rules that were in place in the task, this strategy would not produce success. Instead, a more appropriate strategy to solve the task involved prioritizing important errands, even if this didn't allow participants to follow the shortest route (which was not a strategy participants tended to naturally adopt, even when they were informed their default strategy was not optimal).

The idea that simplistic tasks with reasonably arbitrary rules may lean on idiosyncratic performance factors highlights yet another potential issue in the field - namely that despite being given the same high-level label of “planning task,” they may each actually measure distinct, context dependent skills rather than a single shared coherent construct. This presents a challenge for the field, as these tasks are frequently treated as interchangeable indices of general planning ability despite the dearth of evidence to this effect. Indeed, initial evidence challenging this assumption from our previous work (Cunningham et al., 2025), suggests that performance on several common tasks reported as general planning measures in the literature is not related. We collected participant data on three commonly used psychological planning measurement tasks—the Tower of London, the Traveling Salesperson Problems, and the Zoo Map test. No meaningful correlations were found among the outcome measures, suggesting the absence of a shared cognitive skill underlying performance across these tasks despite each claiming to measure planning as a general function. In fact, the results revealed that performance on each planning task was often more strongly associated with non-planning-related tasks than with the other planning tasks, further negating the assumption that these paradigms measure a unified construct of planning.

Here, we ask whether the inherent characteristics of video games can help bridge the gap in the planning literature between overly simplistic laboratory tasks and highly context-dependent real-world assessments. Specifically, we chose to use mini games. Mini games are particularly well-suited for this purpose because they are designed to be widely accessible: they typically require no prior gaming experience, feature intuitive mechanics, and are freely or inexpensively available online without the need for additional hardware. These features make them ideal for use in research with broad participant samples while preserving engagement and ecological validity. In this study, we sought to leverage these qualities to achieve an intermediate level of task complexity that would capture planning as it unfolded in situ. Following the framework utilized by Baniqued et al. (2013) discussed above, we selected a subset of mini games that demonstrated strong face validity for engaging planning processes. We concentrated on games that required players to think several steps ahead, organize a sequence of actions, and then execute those actions to reach a goal.

As such, the present study examined participants' performance on four planning-related mini games, one non-planning mini game included as a control measure, three commonly used measures of planning, and a number of other individual-difference measures. Although the analyses carried out were largely exploratory, we hypothesized that the planning-based mini games would correlate more strongly with one another than the standard psychological planning measurement tasks correlated with one another—prior work found negligible correlations among the standard planning tasks (Cunningham et al., 2025)—given that they shared key task dimensions such as the length of the planning horizon (or number of visible steps toward the goal state). The identification of such a pattern of bivariate correlations in the present study would be an important first step in this novel line of research, suggesting that the mini games may capture planning-related processes more cohesively and comprehensively than traditional tasks, and underscoring the need for future research utilizing more robust analytic approaches to isolate the specific components or dimensions that engage planning skills within planning-related tasks.

2. Methods

Participants' performance was assessed on a set of five commercially available mini games, four of which were selected for their reliance on planning for success, and on a battery of 14 tasks including cognitive skills metrics, individual difference measures, personality characteristics, and demographic information including video game play experience (see Materials section for details on the games and measures). It is important to note that the current work examines the video game component of a larger study, with a portion of that previous study being reported in Cunningham et al. (2025). Specifically, while correlations between the traditional psychological measures of planning as well as those between the planning measures and other cognitive measures were reported in Cunningham et al. (2025), none of the mini game data was reported previously.

2.1. Participants

A total of 67 participants were included in the final analysis (fully overlapping with the participants included in Cunningham et al. (2025)). No participant's data was fully excluded from the current analysis. However, some data is missing as a result of technical failures (see Supplementary Material Table S1 for details on missing and excluded data points).

All participants were undergraduate students enrolled at the University of Wisconsin-Madison, a large public university in the Midwestern United States. Recruitment occurred through the psychology department's participant pool, and participants received course credit for their involvement. The study procedures were approved by the university's Institutional Review Board for Minimal Risk Research. All participants provided informed consent and were informed of their right to withdraw at any time without penalty. Inclusion criteria required normal or corrected-to-normal vision, no history of major neurological disorders, and an age of 18 years or older. Additional demographic information is provided in Table 1.

Recruitment took place over the course of one semester and concluded once we reached a sample size sufficient for the planned bivariate analyses. This study was designed as an initial investigation into the use of mini games to assess planning-related skills, and the sample size reflected feasibility constraints typical of early-stage exploratory work. Analyses were guided by Bayes Factor thresholds chosen to detect meaningful patterns that could inform future, hypothesis-driven research. In the absence of a definitive benchmark for the strength of correlations that indicate a shared underlying ability, we powered our study to detect associations reflecting approximately 25 % shared variance between measures (Dale et al., 2021).

Table 1
Sample demographic information.

Demographic Category	N	Sample statistics
Age	65	
Mean (SD)		18.6 (1.3)
Min, Max		18.0, 28.0
Gender	65	
Female		35 (54 %)
Male		30 (46 %)
Education completed	65	
High school degree or equivalent		62 (95 %)
Associate degree (2 year college/ junior college)		2 (3.1 %)
Bachelor's degree		1 (1.5 %)
Video game player category[a]	67	
Action video game player (AVGP)		4 (6.0 %)
Tweener		36 (54 %)
Low-tweener		13 (19 %)
Non-action video game player (NVGP)		14 (21 %)

Note: The full sample included 67 participants, however, two participants from the full sample did not provide some demographic information.

[a] See Materials section 2.3 for categorization procedure.

2.2. Study procedure

The study comprised two sessions conducted on separate days. In the first session, participants completed a battery of 13 cognitive and behavioral measures, requiring approximately two hours to complete. In the second session, participants played five online mini video games and completed the Bavelier Lab Video Game Questionnaire (version 2021; Green et al., 2017), which assessed their prior video game experience. The maximum interval allowed between sessions was 10 days. The order of the cognitive measures and the order of the mini games were each counterbalanced: approximately half of the participants completed one sequence of tasks and mini games, while the other half experienced both the task order and the mini game order in reverse. In other words, all participants completed the cognitive tasks on day one of the study and the mini games on day two, but the order of tasks and mini games on each day varied across two possible sequences. This sequencing was intended to mitigate potential attrition by having participants complete the more cognitively demanding and less intrinsically engaging tasks first, with the more enjoyable mini games presented in the second session as an incentive for continued participation—thereby reducing the likelihood of decreased motivation or dropout between sessions.

2.3. Materials

2.3.1. Commercial mini games

The current study included five mini games. Mini games were chosen for this study due to their accessibility and low barrier to entry for novices. They are freely or inexpensively available online, do not require any specialized gaming hardware, and were designed to be approachable for individuals with no prior gaming experience. Participants began at the starting level of each game, with difficulty increasing gradually through the presented levels. Clear instructions on game objectives and controls were provided before play began, and participants were encouraged to ask questions if they encountered confusion during gameplay. Four of the five mini games were selected based on their face validity as planning activities. These games generally required players to look ahead several steps toward the solution state and required a specific sequence of moves to be executed for successful completion of each level (although some games allowed the progression through this necessary sequence to be indirect). Each of these games was played for 12 min and participants were encouraged to progress through as many levels as possible within that time. The final mini game was chosen because it lacked mechanics that allowed players to look ahead or plan sequences of moves in advance, making it unlikely to engage planning skills. This game was included as a control for characteristics inherent in many games that are unrelated to planning (or "gaminess"). Due to time constraints and the overall simplicity of the game, participants only played this game for 8 min.

Game data was collected using the Inputlog research keylogger (Leijten & Van Waes, 2013) installed on a desktop where participants were playing the games in a browser window or through the digital game distribution application *Steam*. Collected data included each individual keystroke, mouse click, movement, scroll, and hover action performed during the recording period. While playing the games, participants were asked to indicate, by typing on the keyboard, the level–or attempt number–they were about to begin. From this data file, for each

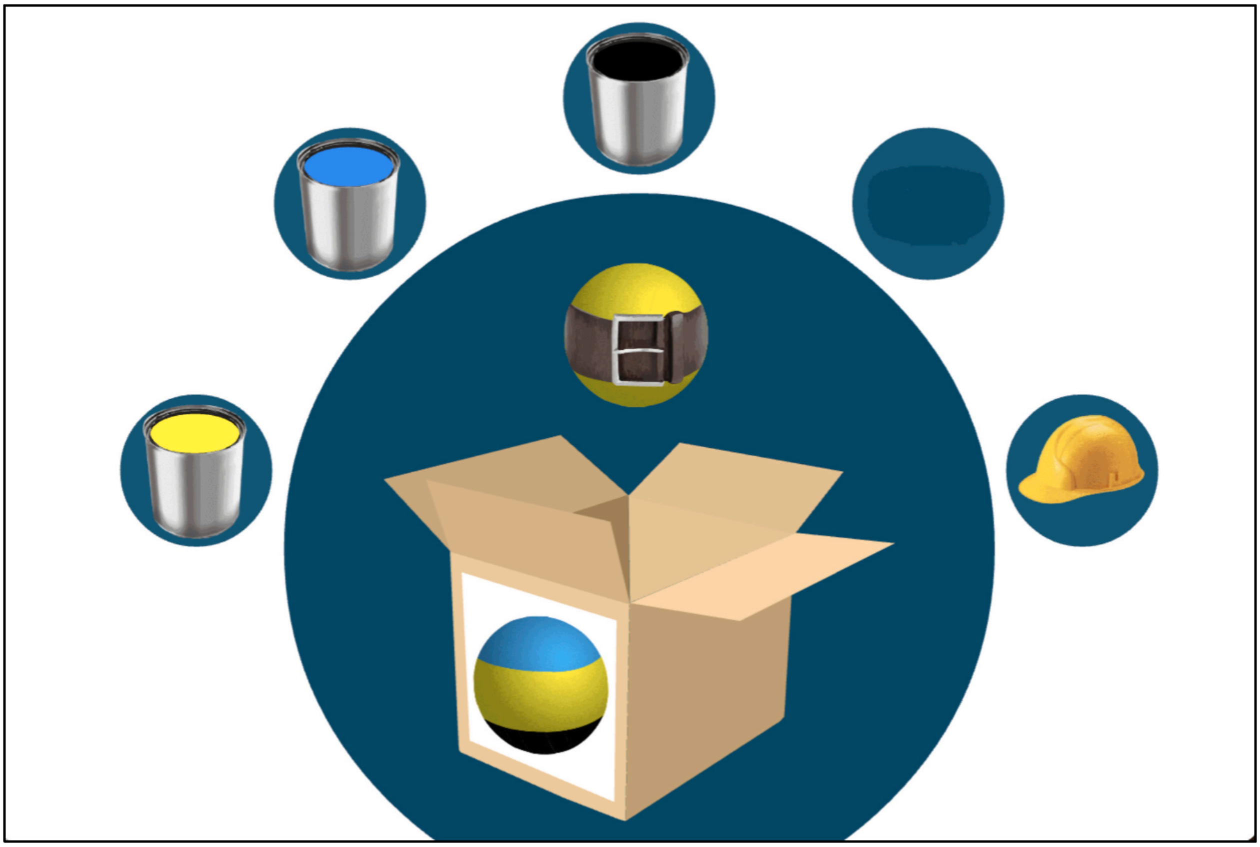


Fig. 1. Factory Balls gameplay example: In Factory Balls, the goal state is given by the ball on the front of the cardboard box. Players start with an initially white ball and must determine the correct sequence of actions to replicate that target paint pattern. In the example above, the initially white ball needs to be dipped in yellow paint (making the whole ball yellow). Next the belt item needs to be applied (this is the current state in the figure image above). After this, the ball needs to be dipped in blue paint. Next, to preserve the yellow center stripe, the player should leave the belt in place and put the construction hat on the ball. Finally, the ball is dipped in black paint to create the black bottom section. Completing the level requires adherence to the exact sequence of steps, otherwise the goal-state pattern will not be reached. See OSF page for gameplay video. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

game, one dependent variable was extracted and analyzed. Further details on the extraction of dependent variables from selected mini games can be found in Section 2.4 Analysis Plan.

2.3.1.1. Mini game descriptions. (See the *Mini Game Gameplay Videos* folder under the *Files* tab in the associated OSF for video clips of gameplay for each game)

(i) In *Factory Balls*, the player is required to use a set of paints and other objects to decorate a white ball to match the indicated target ball characteristics. Because paints/objects are applied to the ball sequentially, players are required to look ahead several steps to the goal state of the ball and plan out how they might sequence their actions to reach this goal state. See Fig. 1 for an example level.

(ii) In *Grabbot*, players control a character tasked with manipulating key objects in each level to move them to their designated positions. Using a set of allowable moves (e.g., the robot/avatar can't turn around within a square, but must instead always move up/down/left/right), players must navigate the game space and interact with objects appropriately to meet the completion criteria. See Fig. 2 for an example level.

(iii) In *Human Resource Machine*, players have to complete a series of tasks by programming the player character's actions using a set of command blocks. Each level requires players to arrange these commands into the correct sequence to accomplish the given goal. See Fig. 3 for an example level.

(iv) In *Robot.Start*, players use arrow keys to navigate a purple cube-shaped robot through the game space. When an arrow key is pressed, the robot moves continuously until it encounters an obstacle. Players must navigate through a series of levels, interacting with blocks and the buttons within the play space to successfully exit the level. See Fig. 4 for an example level.

For each of the four planning mini games, we used the highest level completed by the participant at the end of the 12 min of play as the dependent variable.

In the non-planning game *Music Line*, the player controls the direction that a small square moves on a narrow path. Without intervention, the square moves along a straight path. To modify that straight path, players must press the spacebar. When the spacebar is pressed, the square makes a 90 degree turn. The upcoming path is only revealed a few moments ahead. As such, when the player sees an upcoming turn, they must react quickly and press the spacebar at the appropriate moment in order to turn the square and remain on the path. If a turn is mistimed resulting in the square not staying on the path, the attempt ends, and the player must restart the level (see OSF page for gameplay video). The dependent variable for this game was the highest number of consecutive turns achieved within the 8 min play period.

2.3.2. Cognitive skill measurements

The cognitive battery comprised ten tasks, including three established measures of planning: the Tower of London task (Phillips et al., 2001), the Zoo Map test (Oosterman et al., 2013), and a version of the Traveling Salesperson Problems based on MacGregor and Ormerod (1996). The remaining tasks were drawn from diverse cognitive domains and were included to provide a broader profile of participants' cognitive abilities, allowing for a more comprehensive examination of the skills associated with mini game performance.

2.3.2.1. Planning measures. The Tower of London task is a widely used measure of planning ability in cognitive psychology (Davies, 2005; Phillips et al., 2001). In this task, which is conceptually similar to the Tower of Hanoi, participants manipulate three colored balls across three pegs of varying lengths to replicate a target configuration within a limited number of moves. Each problem allows for up to three attempts. Scoring is based on the number of attempts required to reach the correct solution: three points for solving on the first attempt, two for the second, and one for the third (Krikorian et al., 1994). Higher scores therefore reflect more efficient problem-solving. The dependent variable was the

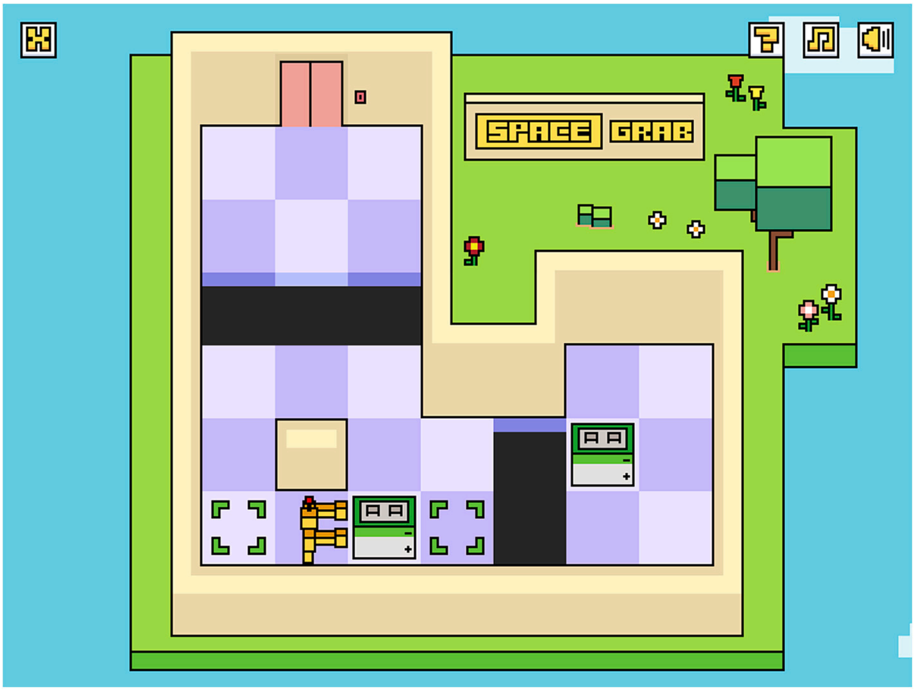
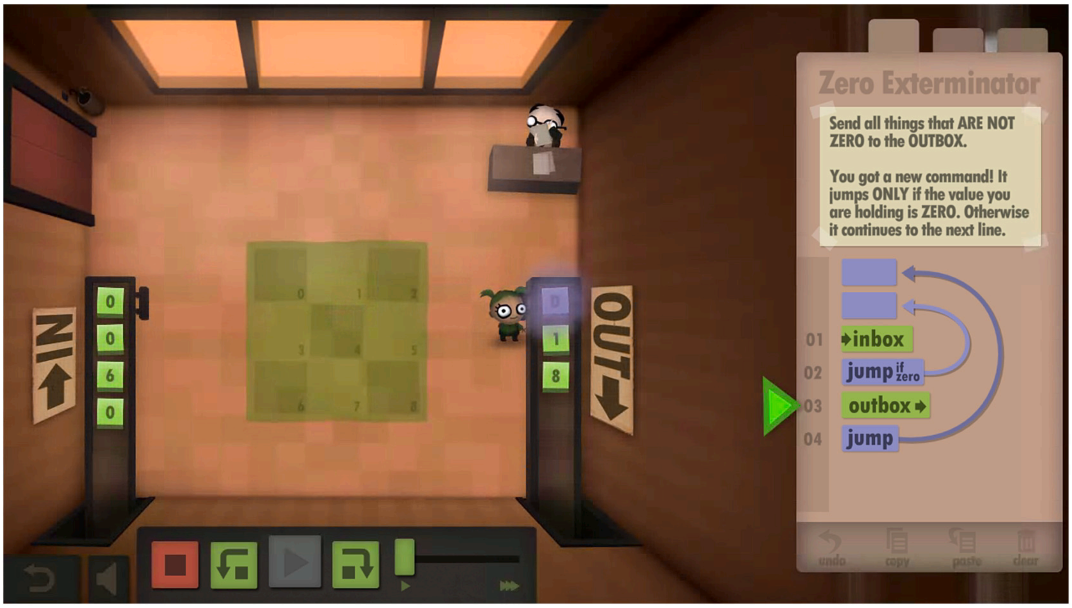


Fig. 2. Grabbot gameplay example: In this level of Grabbot, the goal state is given by the two partially completed squares in the bottom left of the gameplay (checkerboard) area. The two moveable items (battery blocks with the - + symbols on their right hand side) need to be moved to cover the two target squares. To complete this level, the player must navigate across the black chasm to the right side of the play area. Once there, they need to push the green battery block down by one square into its correct position. Afterward, the player must return to the left side and pull the block across the chasm. Success requires careful planning and multiple steps of foresight; without recognizing the sequence of actions needed, the player is likely to encounter repeated failures. See OSF page for gameplay video. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Zero Exterminator

Send all things that ARE NOT ZERO to the OUTBOX.

You got a new command! It jumps ONLY if the value you are holding is ZERO. Otherwise it continues to the next line.

```
inbox
jump if zero
outbox
jump
```

Fig. 3. Human Resource Machine gameplay example: In this game, the objective is to program the avatar to process items from the inbox according to specific rules. For instance, in the level above, the player needs to sequence commands such that all items that are not zero will be sent to the outbox. To complete this level, the player must insert the jump-if-zero command between the inbox and outbox commands in the workspace on the right hand side of the screen. This configuration ensures that when the avatar retrieves an item with a value of zero, it bypasses the action of placing the item in the outbox and proceeds to retrieve the next item. Additionally, the player must place a jump command at the end of the command sequence to enable the process to repeat until all items in the inbox have been handled. See OSF page for gameplay video.

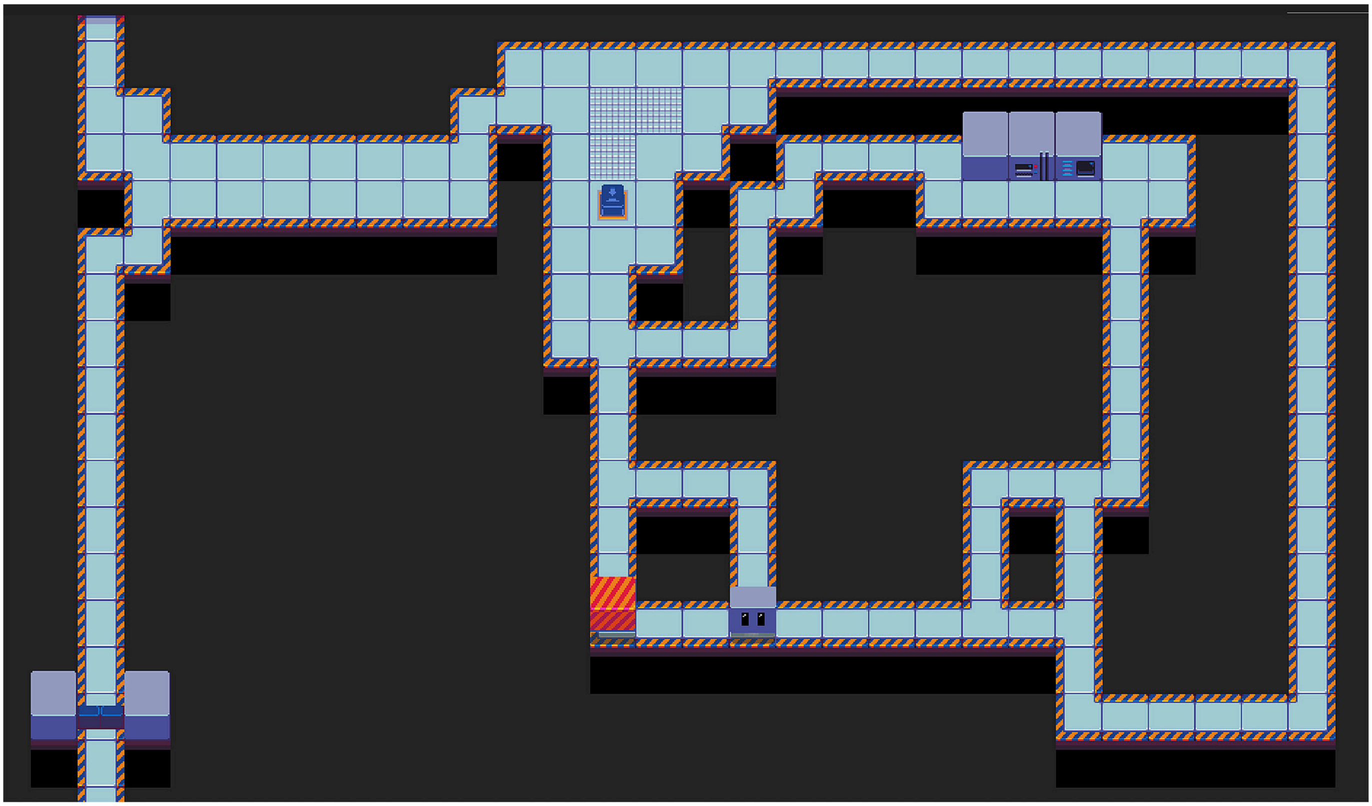


Fig. 4. Robot.Start gameplay example: In Robot.Start, the goal is to move the player's avatar/character (the purple cube-shaped robot) in such a way that they “escape” the level. Pressing a directional key causes the robot to move continuously until it encounters an obstacle. In the level above, the player has just moved to the left and struck the orange block that was previously two spaces to the right. Hitting the orange block stopped the robot just before the upward pathway that is needed to reach the button in the open area above. It also moved the orange block all the way to the left. Given the current state, the robot is free to move up, left, and then up again to press the button in the open area. Moving the robot over this button unlocks the level exit door. The player will then be at the top of the screen above the button. From there they need to move left and down four times to exit (remembering that the robot only stops when it hits a wall or object). See OSF page for gameplay video. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

total score across all 12 problems, with a maximum possible score of 36, following the standard scoring procedure for this version of the task.

The Zoo Map test is a widely used measure of planning ability in both pediatric populations and clinical assessments of adults with neurological disorders, as well as with adult control participants. In this task, participants are required to plan a route through a zoo map in order to visit a predefined set of locations while adhering to specific movement constraints that limit permissible paths (Oosterman et al., 2013). The task includes eight target locations, and there are four valid sequences in which these locations can be visited. The primary dependent variable was the number of locations visited in one of the four correct sequences, consistent with standard scoring practices. Planning time was not included as a measure in this study, as our goal was to align scoring across planning tasks based on accuracy rather than time. Scores ranged from 0 to 8, with a maximum score indicating that all eight locations were visited in a valid sequence.

The Traveling Salesperson Problems, commonly used in computer science to evaluate route optimization in machine learning algorithms, have also been applied in cognitive research to assess human route planning efficiency (MacGregor & Ormerod, 1996). In this task, participants were instructed to draw the shortest possible path that visits each of a set of coordinate points exactly once, without revisiting any point or crossing previously drawn paths. The current study included 14 problems: seven with 10 points and seven with 20 points, each varying in the number of internal nodes to manipulate difficulty (MacGregor & Ormerod, 1996). The primary dependent variable was the total distance traveled between points for each problem, with penalties added for violations of task rules (e.g., skipped points or intersecting paths; see Supplementary Material Table S2 for calculation details). Although prior studies have not typically incorporated distance penalties, doing so was necessary in the present study due to frequent rule-breaking across participants. This adjustment ensured that higher (i.e., less efficient) scores reflected poorer performance, rather than artifacts of incomplete or incorrect task completion. Final performance scores were computed by averaging adjusted distances across all 14 problems. Greater distances indicated less efficient route planning.

2.3.2.2. Fluid intelligence. Fluid intelligence was assessed using the University of California Matrix Reasoning Task (UCMRT). This task requires participants to identify the shape that completes a 3 × 3 matrix pattern composed of various shapes differing in size and color (Pahor et al., 2019). Participants had 10 min to solve as many problems as possible, and their performance was measured as the percentage of correctly solved items out of 23.

2.3.2.3. Spatial cognition. The Mental Paper Folding task measured spatial reasoning and mental imagery. Participants viewed a figure depicting a sequence of folds applied to a square piece of paper, followed by the location of a hole punched in the folded paper. They then selected the correct pattern of holes that would appear when the paper was unfolded, choosing from several options (Shepard & Feng, 1972). Participants completed as many problems as possible within three minutes. The dependent variable was the percentage correct out of 10 problems.

2.3.2.4. Sustained attention. Sustained attention was measured using the gradual-onset continuous performance task (gradCPT). Participants viewed grayscale images of cities (90 % of trials) and mountains (10 %), with each image gradually transitioning to the next over 800 ms via linear pixel interpolation. They were instructed to respond to city scenes with a button press and withhold responses to mountain scenes (Rosenberg et al., 2016). The task consisted of 500 trials. The dependent variable was d-prime, calculated as z(hit rate) minus z(false alarm rate). To avoid calculation issues, hit and false alarm rates of 1 or 0 were replaced with values close to—but not equal to—1 or 0, respectively, as is standard in the literature (Rosenberg et al., 2016).

2.3.2.5. Working memory. Working memory was assessed using the Backward Corsi Block Tapping Span task. Participants viewed eight blocks, with subsets highlighted sequentially, and were required to replicate the sequence in reverse by clicking the blocks in the opposite order (Kessels et al., 2008). Two attempts were allowed at each sequence length, starting from two blocks and increasing until the participant failed twice at a given length. The dependent variable was the maximum sequence length correctly recalled (block span).

2.3.2.6. Cognitive flexibility. The Predictable Task Switching task measured cognitive flexibility by requiring participants to alternate between two binary discrimination tasks—identifying whether a digit was even or odd, or whether a letter was a vowel or consonant—based on the location of stimuli within a 2 × 2 matrix. The target pair moved clockwise through matrix quadrants, and responses were made via key press (Rogers & Monsell, 1995). The dependent variable was the difference in average reaction time between switch trials (where the current task differed from the previous trial) and non-switch trials (where the task repeated). Incorrect and post-error trials were excluded from analysis.

2.3.2.7. Cognitive placekeeping. The UNRAVEL task was employed to measure placekeeping ability during task interruptions. Participants were required to maintain their position within a sequence of binary choice tasks despite interruptions from a secondary, mentally demanding task (Altmann & Trafton, 2015). The dependent variable was the average reaction time on trials immediately following interruptions.

2.3.2.8. Speed of processing. Processing speed was measured via a task requiring participants to quickly identify the direction an arrow was pointing by button press (Dale et al., 2021). The dependent variable was the average reaction time across trials, excluding incorrect and post-error trials.

2.3.3. Additional individual difference measures

2.3.3.1. Crystalized intelligence. Crystallized intelligence was assessed using the Vocabulary subtest of the Shipley Institute of Living Scale. In this task, participants selected the closest synonym from a set of options for 43 common and uncommon words (Shipley, 1940). The dependent variable was the percentage of correct responses out of the 43 items.

2.3.3.2. Creativity/hypothesis generation. Creative thinking was measured with the Shifting Shapes task, in which participants were instructed to create as many novel shapes as possible within five minutes using a set of ten connected squares. Following this, they selected their five most interesting shapes from those generated (Hart et al., 2017). The total number of shapes created within the time limit served as the dependent variable.

2.3.3.3. Personality characteristics. The ten-item version of the Big Five Personality Inventory (BFI) was used to assess personality traits across five domains: agreeableness, conscientiousness, neuroticism, extraversion, and openness. For each trait, a composite score was calculated by combining responses to one forward-coded and one reverse-coded item (Rammstedt & John, 2007). The abbreviated BFI-10 was employed instead of the full 44-item Big Five Inventory to reduce participant burden given the study's length. Prior research has demonstrated that the BFI-10 is a reliable alternative, exhibiting an average correlation of 0.83 with the full scale across three samples (based on Fisher's r-to-Z transformation). Correlations for individual personality dimensions ranged from 0.74 to 0.90 (Rammstedt & John, 2007).

2.3.3.4. Video game play experience. Finally, the Video Game

Questionnaire was included to quantify the video game play experience of participants in the study, adapted from Green et al. (2017). Participants were classified into game player categories based on their questionnaire responses: non-players (NVGP) who played little to no video games, action-game players (AVGP) who played substantial amounts of video games within action game genres, low-action game players with limited overall gaming experience ("Low Tweeners"), and players with general gaming experience who did not fit into any of the other categories ("Tweeners"; see Table 1 for demographic statistics on player category).

2.4. Analysis plan

Prior to data analysis, data cleaning and processing was performed separately on the mini game data and on each of the tasks in the individual difference metric battery. Because the mini game data was collected through a keylogger developed for research purposes (Leijten & Van Waes, 2013), data files needed to be processed significantly in order to extract useful dependent variables, which was done using Python (Python Software Foundation, 2020). To achieve this, the following steps were undertaken to process the raw keylogger data into meaningful dependent variables for use in subsequent analyses: 1) one separate file was saved per individual participant and converted into a manipulatable format via Python, 2) each participant file was cleaned such that only relevant actions were retained, 3) files were subset by game and levels achieved were counted within each game, 4) scores were recorded in a separate file with all participants scores for each game included for further analysis. All subsequent analyses were conducted in R (R Core Team, 2024). Note that for the Music Line game, an additional step occurred between steps 3 and 4 above, where a custom R script calculated the largest number of sequential moves the participant made in a single attempt, which was then recorded as the dependent variable for this game.

As a first step toward investigating the relationships between performance on the selected mini games, we calculated bivariate correlations among the dependent variables collected from the mini games. Additionally, we examined the relationships between the games and other metrics collected in the study to identify potential associations with related skills and traits. This approach served as an initial exploration into the cohesiveness of performance on the planning-related games and their strength of association with planning tasks and with broader cognitive and behavioral constructs. We computed further exploratory analyses in order to better understand the relationships that emerged from the bivariate correlational results. Below we discuss the most interpretable or interesting results, while providing data for the remaining associations not discussed explicitly.

3. Results

Spearman's rank correlations between all dependent variables from the planning-related mini games and the additional measurement tasks were computed to explore the relationship between performance on each individual game or task (Spearman, 1904). For consistency across measures, dependent variables where lower scores indicated better performance were reversed so that higher values uniformly reflected better outcomes across all tasks and measures. Correlations were further evaluated using Bayes Factors to interpret evidence for either the null or alternative hypothesis. With our sample size of 67, we calculated the Bayes Factor test statistic thresholds (null and alternative) using the function *bayesThreshold* from the R package EGCfuncs (v0.1.0, Cunningham, 2025) in alignment with Morey and Rouder (2022). The null acceptance range was $|r| < 0.17$, indicating a likely equivalent correlation to 0, while the null rejection range was calculated as $|r| > 0.31$. Values below $|r| = 0.17$ are used to infer a lack of relationship between measures. Values above the null rejection threshold ($|r| > 0.31$) are considered support of a relationship. Values falling between these threshold values indicate a potential relationship but require further evidence to establish strong confidence.

To exclude general video game experience as a primary explanation for performance on the mini games, we conducted an exploratory analysis examining the correlation between general commercial video game experience and performance on each mini game. To do this, we used participants' self-reported average hours of gameplay across all genres during a typical heavy gaming week in the past year, and separately, prior to the past year, as measured by the Bavelier Lab Video Game Questionnaire (v.2021, Green et al., 2017). Spearman's rank correlations between hours played and each mini game performance outcome were weak or nonexistent for both timeframes, suggesting that performance on the mini games was unlikely to be primarily driven by either recent or long-term commercial video game experience. See Supplementary Material Table S3 for these correlational values.

3.1. Planning mini game performance results

The top section of Table 2 shows the correlations between the planning-related mini games. In support of our hypothesis that the planning-related mini games tap into similar skills, nearly all the games were correlated with one another over the null acceptance threshold (rs > 0.17) except the correlation between the games Grabbot and Human Resource Machine which was just below the threshold ($r = 0.15$). While the remaining correlation values exceeded the null acceptance threshold, they fell short of the alternative hypothesis threshold of $r >$ 0.31, which would have indicated a strong and significant relationship between the performance metrics. One correlation did surpass this threshold: the relationship between the Human Resource Machine game and the Robot.Start game ($r = 0.42$), signifying a reliable correlation between the two. However, these two games also correlated with the control mini game ($r = 0.23$ for Human Resource Machine and $r = 0.38$ for Robot.Start), suggesting that their shared variance may stem from a common reliance on processing speed skills essential for success in the control game. The remaining two mini games, Factory Balls and Grabbot, showed a moderately strong correlation with each other, just reaching but not exceeding the threshold for the alternative hypothesis ($r = 0.31$). Unlike the Human Resource Machine and Robot.Start games, these two games did not correlate with the control game (rs < 0.17), suggesting that their shared variance is likely not attributable to processing speed skills (See Supplementary Materials Fig. S1 A through F for graphical representations of the relationships of interest and discussion above).

3.2. Mini game and planning task performance results

In Cunningham et al. (2025), we demonstrated that performance on the various planning tasks did not correlate. However, while the lack of inter-task correlations suggests the standard planning tasks are not measuring the same construct, it could still be the case that one task could emerge as a more valuable measure than the others and thus correlate with the planning mini games. In order to answer this question, we considered the correlations between the mini games and the traditional planning tasks. As shown in Table 2, generally correlations between the mini games and the planning tasks did not exceed the null hypothesis threshold ($r > 0.17$). However, one value did exceed the alternative hypothesis threshold ($r > 0.31$) which was the correlation between the Traveling Salesperson Problems and the mini game Robot. Start ($r = 0.33$). A lack of robust or consistent correlations between games and tasks suggests that performance on these tasks might be idiosyncratic, or that better performance is due to superior skills in other domains such as speed of processing. This is further supported by the more consistent correlations between the classic tasks and the two mini games which are strongly correlated with the control mini game in which success is predicated on processing speed.

Table 2
Correlations between mini games, planning measures, and other measures.

Measured Construct/Activity (Task Name)	Factory Balls	Grabbot	Human Resource Machine	Robot.Start
Selected Planning-Related Mini Games				
Factory Balls - Levels Passed	X	**0.31**	**0.28**	**0.23**
Grabbot - Levels Passed		X	0.15	**0.22**
Human Resource Machine - Levels Passed			X	**0.42***
Robot.Start - Levels Passed				X
Planning Measures				
Planning (Tower of London)	0.10	0.12	**0.20**	0.11
Planning (Zoo Map)	-0.07	-0.10	0.09	**0.22**
Planning (Traveling Salesperson)	0.09	0.16	0.12	**0.33***
Other Individual Difference Measures				
Control Mini Game (Music Line - Longest Run)	-0.21	0.04	**0.23**	**0.38***
Speed of Processing (Arrows RT)	-0.04	0.14	**0.18**	**0.18**
Placekeeping (UNRAVEL - RT)	0.07	**0.20**	**0.39***	**0.25**
Fluid Intelligence (UCMRT)	**0.19**	**0.20**	0.09	-0.11
Creativity (Shifting Shapes - Total Shapes)	**0.30**	**0.24**	0.11	**0.25**
Spatial Reasoning (Paper Folding)	**0.40***	0.12	**0.26**	0.05
Spatial WM (Backwards Corsi Block Tapping)	0.14	0.05	**0.17**	0.15
Sustained Attention/Inhibition (GradCPT)	0.00	-0.02	**0.24**	0.03
Multitasking (Predictable Task Switching)	0.04	0.14	-0.06	-0.05
Crystalized Intelligence (Shipley Vocabulary)	0.05	-0.02	**0.1**8	0.14
Extraversion (BFI)	0.07	-0.06	0.09	**0.26**
Agreeableness (BFI)	-0.05	0.00	-0.10	**-0.18**
Conscientiousness (BFI)	-0.16	-0.10	0.07	-0.02
Neuroticism (BFI)	-0.08	**-0.32***	-0.11	**-0.28**
Openness (BFI)	0.09	-0.15	0.00	0.08

Note: All correlations are Spearman's rank correlations rounded to the 2nd decimal. Bolded values are greater than the null threshold of 0.17 and starred values are greater than the alternative threshold of 0.31.

3.3. *Mini game and individual difference measure results*

Finally, correlations were calculated between the planning-related mini games and the remaining cognitive and personality metrics collected in the battery of tasks. Notably, correlations emerged between the measure of processing speed and the same two games, Human Resource Machine and Robot.Start, that were correlated with the processing speed related control mini game, Music Line (rs = 0.18). Additional correlations between these same mini games and a measure of reaction speed in the placekeeping task (UNRAVEL; r = 0.39 with Human Resource Machine and r = 0.25 with Robot.Start), are consistent with the general patterns of an overall reliance on processing speed in these games discussed previously. No correlations were found between the speed of processing task and the other two games, Factory Balls and Grabbot. Continuing this pattern, weak or non-inexistent correlations were found between these games and the UNRAVEL task. These results support a pattern in which two of the mini games, Human Resource Machine and Robot.Start, appear to load heavily on processing speed skills while the other two mini games, Factory Balls and Grabbot, do not share this reliance. Several additional measures, which have been associated with planning previously, shared correlational values with Factory Balls and Grabbot, such as fluid intelligence and creativity (rs > 0.17). The strongest correlation between games and tasks was between the mini-game Factory Balls and the paper folding task, a measure of spatial reasoning abilities. This correlation exceeded the alternative hypothesis threshold (r = 0.40), suggesting that the game Factory Balls engages skills similar to those assessed by the paper folding task—namely, spatial reasoning abilities. Several other metrics shared correlational values with the planning-related mini games that were of a magnitude which indicated that they could be interesting to explore in future work (see Table 2).

Correlations between the tasks themselves are not the focus of the presented analyses but see Supplementary Materials Fig. S2 for a graphical representation of these relationships.

4. Discussion

Video games exist at the junction between traditional and often overly simplistic cognitive tasks, and the real-world contexts within which researchers hope to understand the cognitive skills under examination. Past work has identified the need for making measurements at this level and has explored the use of games to assess complex cognitive skills such as spatial reasoning and problem solving in the past with success (Shute et al., 2016; van Opheusden & Ma, 2019). The present work sought to provide the foundation for potentially using what has been called "stealth assessment" to evaluate the complex skill of planning. We approached this by measuring participants' performance on a suite of mini video games that required planning, as defined in the literature. We then explored the bivariate correlations between these games and additional cognitive tasks, as well as a mini game that required no planning skills, which served as a control activity.

4.1. Planning mini games

Our results showed that among the four planning-based mini games, two consistent sets of associations emerged. Two of the planning-based mini games, Robot.Start and Human Resource Machine, correlated strongly with one another. They also both had relatively strong correlations with the reaction time-based mini game and less strongly but crucially, with the speed of processing task and with the reaction time based placekeeping task (UNRAVEL), among others. Together these associations point to these two planning games loading more heavily on processing speed and speeded decision making than on planning itself. These results may illustrate a more general principle in cognitive psychological research in which individuals who can think quickly will be successful at planning in addition to a host of other activities. Similarly, although Robot.Start strongly correlated with one of the existing planning measurement tasks, the Traveling Salesperson problems, this likely indicates that speeded processing skills are capturing more variance in this task than planning abilities as well.

The other two of the four planning-based mini games, on the other hand, appear to arise as possibly effective measures of a less speed-confounded planning construct: Factory Balls and Grabbot. These two games clustered together with other meaningful skills for planning as it is generally understood in the literature. These underlying skills include fluid intelligence and creativity, which could be considered an assessment of hypothesis generation or "out of the box thinking." However, in contrast with Robot.Start and the Human Resource Machine, these two games did not correlate at all with our control non-planning related mini game, nor with our measure of speed of processing. Thus, while these tasks may benefit from faster processing, they do not rely on it as the most representative factor driving performance. This result suggests that the association between these two games was likely not due to their reliance on processing speed or general "gaminess", but instead, due to the necessary skills loaded upon by each of the games. These correlational patterns, including associations with additional cognitive skills likely related to successful planning such as fluid intelligence (i.e., recognizing patterns/adapting plans) and creativity (ability to generate novel hypotheses), are indicative that these two games may isolate higher-order cognitive processes integral to planning. While such overlap might raise concerns about measurement impurity, we argue that this reflects the multifaceted nature of planning itself. As a higher-order function, planning necessarily draws upon more basic cognitive skills. That these mini games engage such skills, while relying less on commonly dominant cognitive abilities like processing speed, supports the view that they capture meaningful components of real-world planning. This added complexity may further offer distinct advantages over traditional planning tasks such as increased ecological validity. Indeed, the lack of convergence among traditional tasks raises the possibility that they capture only narrow subcomponents of planning (e.g., look-ahead or rule-following), rather than the broader construct which might rely more heavily on other underlying skills. While such precision may be useful in isolating specific processes in cases of acute neurological deficits and the like, it limits the ecological validity of these tasks and their measurement of planning as it naturally occurs in everyday contexts.

Although these two games (Factory Balls and Grabbot) may engage similar underlying cognitive skills—namely, planning—they differ significantly in surface features, mechanics, and goals, and may place different emphasis on other skills in addition to planning. Individually these games show interesting correlations with other tasks in the battery that may suggest important characteristics that set these two games apart from one another even while both may primarily rely on planning abilities. For example, Factory Balls was strongly correlated with performance on the spatial reasoning task, Paper Folding. This suggests that Factory Balls relies heavily on a more spatial component of planning which would likely indicate a robust reliance on visually simulating the physical states that the problem space objects would need to progress through in order to reach the solution state. This and other intriguing associations between the included mini games and items in the task battery indicate a need for deeper exploration of the components of planning and the tasks which necessitate it.

This work provides a foundation for selecting the most informative measures for future studies, including longitudinal designs, with mini games particularly well suited for such efforts due to their many predesigned levels of increasing difficulty.

4.2. Future work

In pursuing this new avenue for measuring planning, research should focus on navigating the tension between the experimental control provided by more sterile testing environments on one hand, and the rich and enticing environment that larger scale games provide. From one end, research should seek to continuously increase the complexity of task-based and computationally tractable testing environments (van Opheusden & Ma, 2019), perhaps while, at the same time, attempting to increase intrinsic motivation and engagement through the use of game-design principles. It's critical here though to note that the process of such "gamification" is not trivial (Deterding et al., 2011; Plass et al., 2015). While some promising results have been seen in the gamification space (Cagiltay et al., 2015; Klingberg et al., 2005; Ozcelik et al., 2013; Toups et al., 2009), concerns have been raised as well. For example, in a study by Katz et al. (2014) where popular motivational "gamification" features such as a real time score were added to a traditional cognitive training task, learning was found to be worse than in the original version. The authors suggest that this effect may have been due to the gamification aspect distracting from the core learning component. This highlights the potential risk associated with adding "game-like" motivational components without fully and holistically integrating those components into the experience. A systematic review of cognitive trainings and assessments with such tacked-on components by Lumsden and colleagues elucidates this point further, indicating that the addition of gamification elements to cognitive trainings in general may reduce data quality and/or diminish intervention effects (Lumsden et al., 2016). The distinction between simply "gamifying" an otherwise unmotivating or tedious task and creating a genuinely engaging game experience is critical. Superficial gamification is unlikely to evoke a true sense of fun or playfulness—often described as "chocolate-covered broccoli," where an undesirable task is superficially masked by seemingly desirable elements. This approach contrasts with more effective game-based learning strategies as discussed by Plass et al. (2020), which emphasize playfulness as a core quality of success.

From the other end, research should also continue to seek out ways to utilize or lightly adapt already enjoyable games to aid in cognitive measurements. While, as was seen here, this approach is promising, it also has points requiring caution. Although approaching measurement from a more intrinsically motivating baseline, researchers using entertainment-based games lack fundamental control over the tasks at hand. An obvious initial concern when using commercial games is the issue of consistent availability and functionality of the games of interest. For example, commercial games may not be regularly updated or may cease to be maintained by their creators which poses challenges for researchers using the game long term. In contrast, games with active developers can be changed without notice. Developers may update or change the available levels, mechanics, or even goals and structure of the game at will, thus potentially destroying the usefulness of a game to a particular researcher or team. An additional, and perhaps more often cited, challenge of using commercial video games for research purposes is the lack of access to backend servers. Without access to log files and detailed data about players' actions and game-defined achievements, researchers may struggle to determine and extract dependent variables of interest. For instance, in the case of the current study, even with relatively simple games, we needed to collect data via keystrokes recorded locally on the computer participants used to play the games. Although this data was relatively comprehensive in terms of the actions a player made (e.g., their cursor movements, mouse clicks, and key presses), the dependent variables of interest for each of the games needed to be recorded manually via a specially defined extractable key sequence (in our case the letter “L” and then the numerical digits representing the next attempted level). This posed many problems for accurate data collection and extraction as it depended on some level of participant compliance or direct intervention by an experimenter.

In all, the difficulties associated with gamifying tasks and, simultaneously, the challenges of using commercial games for measurement purposes, could be ameliorated by an increase in partnerships between commercial game developers and psychologists. Some such partnerships exist at present, including the Brain Game Center at Northeastern University and the Neuroscape project at the University of California-San Francisco (UCSF). These research centers, along with other forthcoming projects, can make strides toward achieving this middle ground between simplistic tasks developed for research purposes and complex commercial games developed for engagement from the general public. These endeavors should seek to retain the intrinsically motivating and cognitively demanding components of games for entertainment, while injecting the experimental tractability of traditional psychological measures into activities employed to measure complex cognitive skills, including planning skills.

CRediT authorship contribution statement

Emma G. Cunningham: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Daphné Bavelier:** Writing – review & editing, Methodology, Conceptualization. **C. Shawn Green:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Ethics approval

Approval was obtained from the Minimal Risk Research IRB at the University of Wisconsin-Madison. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Funding

This work was partially supported by a grant from the Office of Naval Research N00014-22-1-2283 to C.S.G. and the Training for Future Learning grant from the Office of Naval Research N00014-20-1-2074 to D.B. Additionally, this work was partially supported by the Klaus J. Jacobs Foundation funding to D.B.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Daphne Bavelier reports financial support was provided by Office of Naval Research N00014–20-1-2074. C. Shawn Green reports financial support was provided by Office of Naval Research N00014–22-1-2283. Additionally, this work was partially supported by the Klaus J. Jacobs Foundation funding to D.B.. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2025.105252.

Data availability

Data and supplementary materials can be accessed via the project OSF site: https://osf.io/fqxev.

References

Altmann, E. M., & Trafton, J. G. (2015). Brief lags in interrupted sequential performance: Evaluating a model and model evaluation method. *International Journal of Human-Computer Studies, 79*, 51–65. https://doi.org/10.1016/j.ijhcs.2014.12.007

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

Bavelier, D., & Green, C. S. (2019). Enhancing attentional control: Lessons from action video games. *Neuron, 104*(1), 147–163. https://doi.org/10.1016/j.neuron.2019.09.031

Bediou, B., Rodgers, M. A., Tipton, E., Mayer, R. E., Green, C. S., & Bavelier, D. (2023). Effects of action video game play on cognitive skills: A meta-analysis. *Technology, Mind, and Behavior, 4*(1). https://doi.org/10.1037/tmb0000102

Bowman, N. D. (2021). Interactivity as demand: Implications for interactive media entertainment. In *The Oxford handbook of entertainment theory* (pp. 647–670). Oxford University Press.

Burgess, P., Simons, J. S., Coates, L. M.-A., & Channon, S. (2005). The search for specific planning processes. In *The cognitive psychology of planning* (pp. 199–227). London, UK: Psychology Press.

Cagiltay, N. E., Ozcelik, E., & Ozcelik, N. S. (2015). The effect of competition on learning in games. *Computers & Education, 87*, 35–41. https://doi.org/10.1016/j.compedu.2015.04.001

Chisholm, J. D., & Kingstone, A. (2015). Action video game players’ visual search advantage extends to biologically relevant stimuli. *Acta Psychologica, 159*, 93–99. https://doi.org/10.1016/j.actpsy.2015.06.001

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*(6), 588–594. https://doi.org/10.1177/0146167298246003

Cunningham, E. G. (2025). *EGCfuncs: Emma G. Cunningham's general analysis and visualization functions* (R package version 0.1.0). https://github.com/emmagcunningham/EGCfuncs.

Cunningham, E. G., Bavelier, D., & Green, C. S. (2025). Rethinking planning metrics: An analysis of common measurements of planning abilities. *Cognition, 263*, Article 106220.

Cunningham, E. G., & Green, C. S. (2023). Cognitive skills acquired from video games. In *Oxford Research Encyclopedia of Communication*. https://doi.org/10.1093/acrefore/9780190228613.013.1468

Dale, G., Cochrane, A., & Green, C. S. (2021). Individual difference predictors of learning and generalization in perceptual learning. *Attention, Perception, & Psychophysics, 83* (5), 2241–2255. https://doi.org/10.3758/s13414-021-02268-3

Dale, G., Kattner, F., Bavelier, D., & Green, C. S. (2020). Cognitive abilities of action video game and role-playing video game players: Data from a massive open online course. *Psychology of Popular Media, 9*(3), 347–358. https://doi.org/10.1037/ppm0000237

Davies, S. P. (2005). Planning and problem solving in well-defined domains. In *, Vol. 35. The cognitive psychology of planning.* Psychology Press. Current issues in thinking & reasoning.

De Schutter, B. (2011). Never too old to play: The appeal of digital games to an older audience. *Games and Culture, 6*(2), 155–170. https://doi.org/10.1177/1555412010364978

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

Eisenberger, R., & Leonard, J. M. (1980). Effects of conceptual task difficulty on generalized persistence. *The American Journal of Psychology, 93*(2), 285–298. https://doi.org/10.2307/1422233

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*(10), 850–855. https://doi.org/10.1111/j.1467-9280.2007.01990.x

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment, 1*(1), 20. https://doi.org/10.1145/950566.950595

Gee, J. P. (2007). *Good video games + good learning: Collected essays on video games, learning, and literacy*. Peter Lang.

Green, C. S., Kattner, F., Eichenbaum, A., Bediou, B., Adams, D. M., Mayer, R. E., & Bavelier, D. (2017). Playing some video games but not others is related to cognitive abilities: a critique of Unsworth et al. (2015). *Psychological Science, 28*(5), 679–682. https://doi.org/10.1177/0956797616644837

Green, C. S., Sugarman, M. A., Medford, K., Klobusicky, E., & Bavelier, D. (2012). The effect of action video game experience on task-switching. *Computers in Human Behavior, 28*(3), 984–994. https://doi.org/10.1016/j.chb.2011.12.020

Hart, Y., Mayo, A. E., Mayo, R., Rozenkrantz, L., Tendler, A., Alon, U., & Noy, L. (2017). Creative foraging: An experimental paradigm for studying exploration and discovery. *PLoS One, 12*(8), Article e0182133. https://doi.org/10.1371/journal.pone.0182133

Hayes-Roth, B. (1980). *Human planning processes* (R-2670-ONR). *The Rand Corporation.* https://www.rand.org/content/dam/rand/pubs/reports/2007/R2670.pdf.

Jones, M. B. (1984). Video games as psychological tests. *Simulations and Games, 15*(2), 131–157. https://doi.org/10.1177/0037550084152001

Katz, B., Jaeggi, S., Buschkuehl, M., Stegman, A., & Shah, P. (2014). Differential effect of motivational features on training improvements in school-based cognitive training. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00242

Kenwright, B. (2023). Data collection during active gameplay: Unveiling a multifaceted portrait of the individual. In *Games as stealth assessments* (pp. 192–219). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0568-3.ch008.

Kessels, R. P. C., van den Berg, E., Ruis, C., & Brands, A. M. A. (2008). The backward span of the Corsi block-tapping task and its association with the WAIS-III digit span. *Assessment, 15*(4), 426–434. https://doi.org/10.1177/1073191108315611

Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What makes a strong team? Using collective intelligence to predict team performance in league of legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 2316–2329). https://doi.org/10.1145/2998181.2998185

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., … Westerberg, H. (2005). Computerized training of working memory in children with ADHD-A randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(2), 177–186. https://doi.org/10.1097/00004583-200502000-00010

Krikorian, R., Bartok, J., & Gay, N. (1994). Tower of London procedure: A standard method and developmental data. *Journal of Clinical and Experimental Neuropsychology, 16*(6), 840–850. https://doi.org/10.1080/01688639408402697

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

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), Article e5888. https://doi.org/10.2196/games.5888

Ma, I., Phaneuf, C. V., van Opheusden, B., Ma, W. J., & Hartley, C. (2022). Distinct developmental trajectories in the cognitive components of complex planning. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9c91h9f7.

Macgregor, J. N., & Ormerod, T. (1996). Human performance on the traveling salesman problem. *Perception & Psychophysics, 58*(4), 527–539. https://doi.org/10.3758/BF03213088

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of Bayes factors for common designs* (version 0.9.12.4.7) [R package]. *Comprehensive R Archive Network (CRAN).* https://CRAN.R-project.org/package=BayesFactor.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). San Diego, CA: Academic Press.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*(6), 1314–1334. https://doi.org/10.1037/a0012702

Oosterman, J. M., Wijers, M., & Kessels, R. P. C. (2013). Planning or something Else? Examining neuropsychological predictors of zoo map performance. *Applied Neuropsychology: Adult, 20*(2), 103–109. https://doi.org/10.1080/09084282.2012.670150

Ozcelik, E., Cagiltay, N. E., & Ozcelik, N. S. (2013). The effect of uncertainty on learning in game-like environments. *Computers & Education, 67*, 12–20. https://doi.org/10.1016/j.compedu.2013.02.009

Pahor, A., Stavropoulos, T., Jaeggi, S. M., & Seitz, A. R. (2019). Validation of a matrix reasoning task for mobile devices. *Behavior Research Methods, 51*(5), 2256–2267. https://doi.org/10.3758/s13428-018-1152-2

Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the tower of London task. *The Quarterly Journal of Experimental Psychology Section A, 54*(2), 579–597. https://doi.org/10.1080/713755977

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Plass, J. L., Homer, B. D., Mayer, R. E., & Kinzer, C. K. (2020). Theoretical foundations of game-based and playful learning. In *Handbook of game-based learning* (pp. 3–24). The MIT Press.

Python Software Foundation. (2020). Python (Version 3.9) [Programming language]. https://www.python.org/.

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing https://www.R-project.org/.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Rosas, R., Ceric, F., Aparicio, A., Arango, P., Arroyo, R., Benavente, C., … Véliz, S. (2015). ¿Pruebas Tradicionales o Evaluación Invisible a Través del Juego?: Nuevas Fronteras de la Evaluación Cognitiva. *Psykhe (Santiago), 24*(1), 1–11. https://doi.org/10.7764/psykhe.23.2.724

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., & Chun, M. M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience, 19*(1), 165–171. https://doi.org/10.1038/nn.4179

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain, 114*(2), 727–741. https://doi.org/10.1093/brain/114.2.727

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology, 3*(2), 228–243. https://doi.org/10.1016/0010-0285(72)90005-9

Shipley, W. C. (1940). *Shipley Institute of Living Scale (SILS)*. APA PsycTests.

Shute, V., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior, 116*, Article 106647. https://doi.org/10.1016/j.chb.2020.106647

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

Shute, V., & Wang, L. (2015). Measuring problem solving skills in portal 2. In P. Isaías, J. M. Spector, D. Ifenthaler, & D. G. Sampson (Eds.), *E-learning systems, environments and approaches: Theory and implementation* (pp. 11–24). Springer International Publishing. https://doi.org/10.1007/978-3-319-05825-2_2.

Shute, V., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Stephen, C., & Edwards, S. (2017). *Young Children Playing and Learning in a Digital Age: A Cultural and Critical Perspective*. Routledge. https://doi.org/10.4324/9781315623092

Tenorio Delgado, M., Arango Uribe, P., Aparicio Alonso, A., & Rosas Díaz, R. (2016). TENI: A comprehensive battery for cognitive assessment based on games and technology. *Child Neuropsychology, 22*(3), 276–291. https://doi.org/10.1080/09297049.2014.977241

Toups, Z. O., Kerne, A., & Hamilton, W. (2009). *Motivating play through score* (ACM Computer Human Interaction Workshop on Engagement by Design).

van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences, 29*, 127–133. https://doi.org/10.1016/j.cobeha.2019.07.002

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

Ventura, M., Shute, V., Wright, T. J., & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00852

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education, 60*(1), 52–58. https://doi.org/10.1016/j.compedu.2012.07.003

Zhang, R.-Y., Chopin, A., Shibata, K., Lu, Z.-L., Jaeggi, S. M., Buschkuehl, M., … Bavelier, D. (2021). Action video game play facilitates "learning to learn.". *Communications Biology, 4*(1), Article 1. https://doi.org/10.1038/s42003-021-02652-7