UCLA CRESST NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

Lyons ASSESSMENT CONSULTING

MODELING THE RELATIONSHIPS AMONG ONLINE SOLITAIRE GAMEPLAY AND MEASURES OF COGNITION

Sam Ihlenfeldt, Gregory K. W. K. Chung, Susan Lyons, Jordan Lawson, and Elizabeth J. K. H. Redman

APRIL 2025

CRESST REPORT **877**

This report was prepared collaboratively across the teams at Lyons Assessment Consulting and UCLA CRESST. This research is supported with funding from Unwind Media, LLC.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of Unwind Media, LLC.

To cite from this report, please use the following as your APA 7th edition reference: Ihlenfeldt, S., Chung, G. K. W. K., Lyons, S., Lawson, J., & Redman, E. J. K. H. (2025). *Modeling the relationships among online Solitaire gameplay and measures of cognition* (CRESST Report 877). UCLA/CRESST.

Table of Contents

Modeling the Relationships Among Online Solitaire Gameplay and Measures of Cognition

Sam Ihlenfeldt,[a] Gregory K. W. K. Chung,[b] Susan Lyons,[a] Jordan Lawson,[a]
and Elizabeth J. K. H. Redman [b]

[a] Lyons Assessment Consulting
[b] CRESST/University of California, Los Angeles

Executive Summary

In this technical document, we investigate the potential of using Solitaired.com as a tool for cognitive assessment. Solitaire can sustain motivation and engage cognitive processes relevant to constructs like mild cognitive impairment (MCI). Solitaire gameplay also addresses many of the limitations of traditional neuropsychological evaluations, such as their unnatural format and lack of inclusivity for low-literacy individuals. In this document, we explore the research by Gielis and colleagues, which reinforces Solitaired as a feasible platform for measuring MCI and identifies game performance metrics that are sensitive to players' cognitive differences.

An initial pilot study using Solitaired.com data examined the associations among gameplay metrics and MCI. Players self-reported their cognitive status, and metrics such as game completion time and average move time were analyzed. Significant correlations aligned with earlier findings by Gielis and colleagues, supporting the viability of Solitaired.com as a cognitive assessment tool. The study then extended this approach with a larger sample, employing a random group design and regression modeling to predict aspects of mental acuity, as measured by TestMyBrain (TMB) testlets from The Many Brains Project.

This study shows that Solitaired.com gameplay variables are statistically and strongly related to working memory, processing speed, and visual short-term memory, as measured by TMB tests. Based on players' interactions with the game, we can predict players' mental acuity scores on three validated assessments of cognition:

1. Working memory score (Flicker Change Detection)
2. Processing speed score (Choice Reaction Time)
3. Visual short-term memory score (Digit Symbol Matching)

A comparison of each of those scales suggests a high degree of overlap, so presenting players' scores from all three models is not advised. Ultimately, we recommend presenting players with an overall percentile score from the Digit Symbol Matching model and a percentile

score that compares players to all other players within their age bracket. Because age is such a meaningful variable in the model, older players will, on average, have lower mental acuity scores. Consequently, allowing those players to compare themselves to others in their age range could be beneficial.

Modeling the Relationships Among Online Solitaire Gameplay and Measures of Cognition

Sam Ihlenfeldt,[a] Gregory K. W. K. Chung,[b] Susan Lyons,[a] Jordan Lawson[a]
and Elizabeth J. K. H. Redman [b]

[a] Lyons Assessment Consulting
[b] CRESST/University of California, Los Angeles

Abstract: In this evaluation study, we investigated the extent to which Solitaired.com's online game, Solitaire, could be used to model players' performance on several validated cognitive tests commonly associated with mental acuity (i.e., memory and processing speed). Prior research found that Solitaire gameplay is affected by mild cognitive impairment and presumably closely related to mental acuity. Thus, we investigated the relationship between measures of mental acuity and Solitaire gameplay on Solitaired.com. Gameplay and self-reported data from players who opted into the evaluation were used to model players' performance on three brief online tests: (a) Choice reaction time, which involves processing speed, response selection/inhibition, and attention (n = 555, R^2 = .53); (b) Digit symbol matching reaction time, which involves processing speed and visual short-term memory (n = 707, R^2 = .54); and (c) Flicker change detection, which involves visual search, change detection, and visual working memory (n = 568, R^2 = .49). The important gameplay variables were mean time per move and use of hints, and the important player background variable was self-reported age. A major implication is how to report the model output information to players. As an engaging game, Solitaire can sustain motivation and elicit many important cognitive processes. Making full use of the information carried in players' interactions in online games—especially those games with a global audience—may provide new opportunities for exploring novel ways to measure cognitive processes in an aging population and, ultimately, to help players better understand their own gameplay performance.

Introduction

There is growing interest in the use of digital games in the cognitive health field. Well-designed games are entertaining and, importantly, elicit sustained effort from players. Digital games can be administered at scale, can be instrumented to record players' actions within the game, and with sufficient validity evidence, may address many of the shortcomings of traditional approaches to neuropsychological evaluation (e.g., face-to-face testing; perceived intrusiveness and unnatural format; little or no relationship to daily living activities; not validated for low-education or illiterate players) (Groznik & Sadikov, 2019; Valladares-Rodríguez et al., 2016).

One key requirement for a game to be considered a candidate for cognitive assessment is its ability to maintain players' motivation—sustained and effortful gameplay over time. The importance of repeated play is that repeated observations can be made, resulting in a more reliable estimate than a single measurement. In addition, measurements taken over time can show a player's progress over months or even years. Another key requirement is that the game requires players to use the mental processes associated with the assessed construct. For example, if a game purports to assess cognitive impairment, then successful gameplay should require high mental acuity related to processing speed in general and attentional, psychomotor, visual, and memory processes in particular.

Mild cognitive impairment (MCI) is a cognitive impairment that does not meet the criteria for dementia, with a deficit in cognition in at least one domain and no functional impairment of daily living activities (Tangalos & Petersen, 2018). Common screening tests to detect MCI are the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005) and the Mini-Mental State Examination (MMSE) (Tombaugh & McIntyre, 1992). These tests measure a person's language, visual skills, memory, orientation, attention, and executive functions (Pinto et al., 2019).

Review of Literature

Using Solitaire as a Potential Platform to Measure MCI

Klondike Solitaire appears to require many mental processes associated with MCI (Gielis et al., 2017). Gielis et al. asked three subject matter experts to rate how strongly various elements of Solitaire gameplay were related to 10 cognitive functions measured by MoCA, MMSE, and another test. The four cognitive functions rated as most prevalent in Solitaire were attention, executive function, object recognition, and abstraction and memory. All three subject matter experts agreed that Solitaire could be used to measure MCI. The experts also noted that processing speed is another important indicator.

Solitaire is popular among older adults, and an online version of Solitaire has been investigated extensively by Gielis and colleagues (Gielis et al., 2017; Gielis, Vanden Abeele, Croon, et al., 2021; Gielis, Vanden Abeele, Verbert, et al., 2021); see also Gielis, 2019a, 2019b; Gielis et al., 2019) as a means to differentiate between older adults with MCI and healthy older adults. These studies are part of a more extensive research base examining the use of video games for cognitive assessment and cognitive training for MCI (e.g., Boot et al., 2008; Groznik & Sadikov, 2019; Pedersen et al., 2023).

In one study, Gielis, Vanden Abeele, Verbert, et al. (2021) showed that players' gameplay performance in Solitaire differed between a healthy sample and a sample diagnosed with MCI. Appendix A reproduces the data reported in Gielis, Vanden Abeele, Verbert, et al. (2021), and we also calculated effect sizes from these data to help determine which indicators are most sensitive to the sample differences. Appendix B contains the definitions of each indicator (Gielis, Vanden Abeele, Verbert, et al., 2021, p. 46).

Gielis, Vanden Abeele, Verbert, et al. (2021) divided the game mechanics into four major categories: Result-based, performance-based, time-based, and execution-based. Result-based indicators reflect overall performance by the end of the game. Performance-based indicators reflect performance during the game. Time-based indicators are related to time, and execution-based indicators are associated with the physical execution of moves. Overall, Gielis, Vanden Abeele, Verbert, et al.'s data suggest the following:

- In general, compared to the healthy sample, the MCI sample had lower performance, spent more time on moves, and had lower (physical) accuracy. The variation in scores (standard deviation) was also higher in the MCI sample.
- The result-based indicators (total score, no. of games solved, overall game time, and total moves) differentiate healthy from MCI samples. The number of games solved and total game time appear to be the most sensitive indicators.
- The finer grained performance-based indicators are less sensitive to detecting differences between the two samples. Except for the *successful move percentage* (*d* = 0.68), all performance-based indicators have smaller effect sizes than the result-based *score* (*d* = 0.64).
- Overall, time-based indicators differentiate the two samples better than performance-based indicators. The MCI sample was markedly slower than the healthy sample on *minimum think time* (*d* = 2.83), *average think time* (*d* = 1.30), *average total time* (*d* = 1.10), and *average move time* (*d* = 1.01). Implications: Solitaired.com provides players' *total moves*, which includes every game action aside from requesting a hint; this can then be averaged across *total time* to yield an *average move time*.
- Overall, execution-based indicators related to accuracy appear to differentiate the two samples. The MCI sample was markedly less accurate than the healthy sample, with *average accuracy* (*d* = 1.05) and *maximum accuracy* (*d* = 0.98) having the largest effect sizes.
- *Hints* and *Undos* were not found to be significant in Gielis, Vanden Abeele, Croon, et al. (2021). However, Wallace et al. (2014) provided a convincing argument that hints were a necessary game feature for adults with MCI.

Gielis, Vanden Abeele, Verbert, et al. (2021) suggest that gameplay in Solitaire is sensitive to MCI. The theoretical account for these results is that diminished mental acuity in the MCI sample manifests in slower processing, diminished working memory, attention, and object recognition (visual processing) (Gielis et al., 2017).

Pilot Study

Viability of Solitaired as a Platform to Measure Mental Acuity Factors Related to MCI

Given the prior research by Gielis and colleagues on using Solitaire to detect MCI, we conducted a pilot study on a convenience sample of Solitaired players. We wanted to examine the extent to which the Solitaired sample performed similarly to what was reported in Gielis, Vanden Abeele, Verbert, et al. (2021). We reasoned that if we observed the same pattern of results as Gielis, Vanden Abeele, Verbert, et al., then that finding would establish an empirical link to prior research and be compelling evidence that gameplay in Solitaired is likely to be sensitive to the various cognitive functions associated with MCI (i.e., mental acuity), but not necessarily MCI itself. Any statistically significant difference would be even more remarkable because of the measurement error due to self-reporting of cognitive impairment (vs. a clinical diagnosis), setting and conditions (uncontrolled vs. controlled), and non-tuned gameplay measures (available gameplay measures vs. optimized measures).

Method

Players' responses to the question, "Do you have a cognitive impairment such as the following: Alzheimer's disease, traumatic brain injury, developmental disability, memory loss?" were used to form two groups (with and without cognitive impairment), and the two groups were compared on game score, game completion time, number of moves, and mean time per move. These measures were chosen because they (a) were close in definition to Gielis, Vanden Abeele, Verbert, et al.'s (2021) result-based measures and time-based measures, and (b) could be computed with the existing Solitaired telemetry. In addition to self-reported cognitive impairment, players were also asked to select an age band that included their age, and the effect of age was examined for each measure. See Appendix C for the full questionnaire.

Results

To understand the viability of a research study aimed at generating a predictive equation for players' mental acuity, we first investigated whether associations existed between gameplay data and players' cognitive ability via exploratory data analyses. Using data provided by Unwind Media on players' *game score*, *game completion time*, *number of moves*, demographic variables such as *age*, and a self-reported measure of cognitive impairment, we examined whether there were group differences between players who reported mild cognitive impairment (or not) on the various gameplay variables. Similarly, we examined whether there were group differences among players by different age bands on the various gameplay variables.

Analyses demonstrated that players' self-reported age was associated with g*ame score*, *game completion time*, and *number of moves*, with older players consistently scoring significantly lower on these outcomes than younger players. Additionally, self-reported cognitive impairment was significantly associated with overall *game score* and *game completion time*, with those reporting being cognitively impaired scoring significantly lower than players reporting no cognitive impairment. These preliminary findings suggested that it would be worthwhile to move forward with the development of a mathematical equation for predicting cognitive impairment from players' gameplay data.

We also compared the pilot study results to the general findings from Gielis, Vanden Abeele, Verbert, et al. (2021). As seen in Table 1, the pilot study results were consistent with Gielis, Vanden Abeele, Verbert, et al.'s results on three of the four gameplay measures. The directions were similar, although the magnitude of the difference (effect size) was lower in the Solitaired sample.

Table 1

Summary of Pilot Study Results and Comparison with Gielis, Vanden Abeele, Verbert, et al. (2021) Results

	Pilot study [a]		
Gameplay measure	By self-reported cognitive impairment status	By age	Comparison to Gielis, Vanden Abeele, Verbert, et al. results
Game score	The impaired group performed lower than the non-impaired group (*d* = 0.22).	Age is associated with *game score*, with performance decreasing with age.	The pilot study results are consistent with those of Gielis, Vanden Abeele, Verbert, et al. *Game score* was lower in the MCI group (*d* = 0.64). Note: The computation for *game score* is likely to differ between Gielis, Vanden Abeele, Verbert, et al. and Solitaired.
Game completion time	The impaired group took longer to complete the game than the non-impaired group (*d* = 0.21).	Age is associated with *game completion time,* with game completion time increasing with age.	The pilot study results are consistent with those of Gielis, Vanden Abeele, Verbert, et al. *Game completion time* was longer in the MCI group (*d* = 0.84).
Number of moves	No difference between groups.	Age is associated with the *number of moves*, with younger players using fewer moves than older players.	The pilot study results are not consistent with the results of Gielis, Vanden Abeele, Verbert, et al. *Number of moves* in the game was higher in the MCI group (*d* = 0.17). Note: The computation for *number of moves* in Gielis, Vanden Abeele, Verbert, et al. may differ from *number of moves* in Solitaired.
Mean time per move	On average, the impaired group took longer to move than the non-impaired group (*d* = 0.19).	Age is associated with the *mean time per move*, with younger players using fewer moves than older players.	The pilot study results are consistent with the results of Gielis, Vanden Abeele, Verbert, et al. *Mean time per move* was longer in the MCI group (*d* = 1.01).

[a] All comparisons are statistically significant at the .05 level unless otherwise noted.

Main Study

Given the associations found within the Solitaired sample and the replication of findings from an existing research base focused on Klondike Solitaire, we concluded that it was plausible to develop and validate a mental acuity scoring procedure for Solitaired's use.

Research Questions

This research seeks to calculate and present players with a mental acuity score using player gameplay data extracted from the game of the day (GoTD) on Solitaired.com. Mental acuity, as used in this research, was measured using cognitive TestMyBrain (TMB) testlets from The Many Brains Project[1] and is defined as processing speed, visual search, and change detection (Passell et al., 2019; Singh et al., 2021). In this report, the following research questions were addressed:

1. What Solitaire gameplay variables predict different facets of mental acuity, after accounting for Solitaire hand difficulty?
    a. Which facet of mental acuity is best explained by a combination of Solitaire gameplay variables?
2. What combination of Solitaire gameplay variables best predicts a single unified mental acuity score?

Method

Design

The study uses a random groups design to investigate associations between players' gameplay behaviors and their mental acuity. Players are randomly assigned to one of five TMB tests. The total number of Solitaired.com players eligible for inclusion was 51,254, although only 33,410 of those players were invited to participate in the study. The total number of players included in the analysis was 4,024. We dropped 551 players from the analysis for the following reasons: (a) extreme Solitaired gameplay time (313 players) and (b) extreme TMB response time (238 players). Sensitivity analyses indicated no substantial impact on the results.

Sampling

The Solitaired.com population comprises international players who access Solitaired.com throughout the day, although most players are from the United States and access to Solitaired.com is nonuniform. Furthermore, the GoTD difficulty varies daily (see Appendix D for GoTD *game difficulty*, as measured by daily win percentages).

[1] https://www.manybrains.net/

Thus, our sampling procedure attempted to mimic these distributional characteristics in two ways. First, to address the variation in GoTD difficulty, data were collected over 88 days and capped at 100 players daily to meet the target of 5000 cases. To address the continuous (but nonuniform) access to Solitaired.com and the international audience, we sampled players across the entire 24 hours of each day. We assumed that if we accepted every eligible player serially, we would reach the 100-player limit well before the end of the 24-hour cycle, resulting in a biased sample. Because the distribution of players throughout 24 hours was nonuniform, a proportion of players was sampled each hour. We expected this sampling strategy to mirror the actual distributional shape. Given our target of 100 players a day, we set the initial sampling proportion to 5%; however, this number was monitored and dynamically increased to ensure the sample size was met. This sampling strategy appeared effective, and no hour of the day had more than 7% of the total sample.

For a player to be included in the study, the following criteria had to be met:

- There were less than or equal to 100 complete players per day (i.e., wins the GoTD and completes the TMB test).[2]
- The player wins the GoTD.
- The player has not participated in the study.
- The player has not previously declined to participate in the study.

Initially, all participating players were also required to register with Solitaired.com. However, upon inspecting the rate of participation, the study team opened the data collection to all Solitaired players. Overall, only 18% of the responses came from nonregistered players. See Appendix E for more information about the inclusion criteria.

Power Analysis

A series of power analyses were conducted to determine the minimum number of Solitaired.com players we would need to respond to each of the cognitive assessment scales developed by TMB to detect effects for our gameplay variables with an 80% chance, given a true effect exists. We ran a set of simulation studies investigating power for the nonmultilevel linear regression setting.

For our power analyses, we investigated what would happen to statistical power (i.e., correctly rejecting the null hypothesis of no effect for our gameplay variables) when sample size and effect sizes for gameplay (i.e., our regression coefficients) were varied for both the multilevel and nonmultilevel regression setting. Power simulations were run using the *simr* and *pwrss* packages in R (Green & MacLeod, 2016).

[2] The initial count was the sum of the number of players who completed the study and the number of players who declined. The number of players who declined was dropped about a month after data collection started.

Power simulations were conducted using nonmultilevel linear regression models consisting of five predictor variables. We examined *R*-squared values (i.e., effect sizes) from .05 to .70 in increments of .05 with alpha set to .05. The simulation results suggested that samples of 200 to 300 players would be adequate for detecting statistically significant *R*-squared values (i.e., *R*-squared > 0) and would also suffice in detecting changes in *R*-squared when employing a hierarchical regression modeling approach. Moreover, given the possibility of adding or controlling for additional player variables and examining for interaction effects and the large sample available for this study, we aimed for a sample size of 1,000 players per scale. This larger estimate helps to ensure adequate power and allows for flexibility in our modeling approach.

Measures

Three kinds of measures were collected from players: (a) self-reported background information, (b) gameplay measures, and (c) mental acuity measures.

Background Measures

Players were asked for the following information: age, sex assigned at birth, and whether they have a cognitive impairment (e.g., Alzheimer’s disease, traumatic brain injury, developmental disability, or memory loss). See Appendix F (Player prompt 2) for the actual screenshot of the question.

Gameplay Measures

Table 2 summarizes the gameplay measures included in the statistical modeling.

Table 2

Gameplay Measures

Gameplay variable	Definition	Units, data type, and possible range
Game completion time	The total time taken for the player to complete the game	Milliseconds, integer (0, +∞)
Number of moves	Every game action, aside from requesting a hint	Count, integer (0, +∞)
Mean time per move	Game completion time / Number of moves	Milliseconds, real (0.0, +∞)
Hint count	Number of hints requested	Count, integer (0, +∞)
Undo count	Number of undos requested	Count, integer (0, +∞)
Hotkey count	Number of card moves were made using a hotkey	Count, integer (0, +∞)
Unproductive moves	Number of unproductive moves, computed as the sum of (a) attempting to place a card that is not a legal move, (b) releasing a card off the canvas, and (c) dragging a card back to its original position	Count, integer (0, +∞)
Game difficulty	For a given hand, the proportion of players who succeed to the total number of players who attempted to solve the hand	Proportion, real (0.0, 1.0)

Note. The inclusion of difficulty in the statistical modeling ensures that strong players are not given a lower mental acuity score solely due to a more challenging hand. While there are many ways to calculate hand difficulty for Klondike Solitaire (Blake, 2020), for measurement purposes, the most effective estimate of difficulty is the proportion of players who beat the hand.

Mental Acuity Measures

Table 3 summarizes the cognitive assessments from TMB that were used in the study. Each of these tests has undergone rigorous psychometric research with large norming samples (The Many Brains Project, 2024).

Table 3

Summary Description of TMB Tests

Test name	Administration time (sec)	Test prompt and cognitive processes measured	Norming sample size [a]	Reliability [b]	Strength of validity evidence [a, c]
Ultra-brief TMB Digit Symbol Matching [a]	90	Using a symbol-number key shown on screen, match as many symbols and numbers as possible in 90 seconds. This test measures processing speed and visual short-term memory.	45,295	.93	S
Ultra-brief TMB Choice Reaction Time [a]	60	Indicate the direction of the arrow that is a different color from the rest. This test measures processing speed, response selection/inhibition, and attention.	18,556	.95	M
Ultra-brief TMB Simple Reaction Time [a]	60	Press a key whenever a green square appears. This test measures basic psychomotor response speed.	49,001	.93	S
TMB Matrix Reasoning [a]	180	Identify the image that best completes the pattern in a series, based on a logical rule. This test measures basic fluid cognitive ability and nonverbal reasoning.	20,510	.89	S
Ultra-brief TMB Flicker Change Detection [a, d]	60	Given a set of flashing blue and yellow dots, find the dot that is changing color from blue to yellow. This is a test of visual search, change detection, and visual working memory.	29,627	.78	S

Note. From the results of Passell et al. (2019), it is not clear if the reliability and validity evidence were collected for the ultra-brief versions of the test.

[a] Passell et al. (2019). [b] Spearman-Brown corrected split-half reliability. [c] S = strong, M = medium. [d] Not smartphone compatible.

Procedure

Pre-study Procedures

In preparation for data collection, Solitaired.com updated their software to meet study requirements depicted in Appendix F and Appendix G. Solitaired.com and TMB worked together to design the interserver communication protocol. System testing focused on verifying data logging for new measures (i.e., number of hints, number of undos, number of hotkey moves, and number of unproductive moves) and verifying operation and data logging on different devices (i.e., Windows and Mac desktop machines, tablets, and Apple and Android mobile devices).

Study Procedures

The study was designed to include players who beat the GoTD and completed the TMB test. Appendix G contains player experience flow, decision points, prompts, and associated screenshots.

The study protocol began when the player won the game of the day. If the player met the inclusion criteria (Appendix E), the player was invited to participate in the study. If the player accepted, they filled out a short questionnaire and completed a TMB test (i.e., a cognitive skills test) delivered by the TMB website. Players who used a nonmobile device were randomly assigned one of five TMB tests. If the player used a mobile device, then the set of TMB tests only included four tests. TMB does not recommend using the TMB Flicker Change Detection test on mobile devices because of the smaller screen size. After completing the TMB test, the player was returned to the Solitaired.com home page.

The protocol was designed to give the player an opt-out option at every decision point. In addition, no personally identifiable information was recorded, and the study data do not contain any information that can be used to connect the data to a particular player on Solitaired.com. To see the actual prompts players received, see Appendix F. To see a description and representative screenshot of the cognitive skills test administered by TMB, see Appendix H.

Finally, during data collection, the data were monitored initially every day, starting on Week 4, every 3 to 4 days. TMB data monitoring included daily TMB tests and distribution of TMB tests throughout the day. Solitaired.com data monitoring included ID uniqueness, daily completes, the number of registered and unregistered players, and variability of gameplay data and player survey data. For more detailed information on the data monitoring protocol, see Appendix I. Table 4 summarizes the major activities and adjustments made to the protocol during data collection.

Table 4

Data Collection Activities

Date	Activity
2024-06-24	Solitaired deployed to 1% of their audience. Players were prompted to participate in the study if the number of completions + number of declines <= 100.
2024-07-24	Data quality confirmed.
2024-07-26	Data collection is fully scaled.
2024-07-26	Inclusion criteria updated. Players were prompted to participate in the study if the number of completions <= 100. The number of declines was no longer considered.
2024-08-10	Inclusion criteria updated for guest players. Guest players were now allowed to register and take the survey. Guest player recruitment scaled at 5%.
2024-08-21	Guest player recruitment was fully scaled.
2024-09-23	Data collection completed.

Analysis

Regression

A series of regression models were fit to each TMB scale in order to determine which gameplay variables were most strongly associated with different facets of mental acuity (as measured by the TMB scales).

$$\begin{aligned} TMB_{ik} = \beta_0 + \beta_1 (Total\,Time) + \beta_2 (Avg.Time/Move) + \beta_3 (Hints) \\ \mathrm{s}{+}\beta_4(Undos) + \beta_5(Gender) + \beta_6(Age) + \beta_7 (Difficulty) + \epsilon_i \end{aligned} \tag{1}$$

where $\boldsymbol{TMB_{ik}}$ is individual i's result from TMB test k. We will be estimating coefficients $\boldsymbol{\beta_0}, \ldots, \boldsymbol{\beta_7}$, which represent the coefficients for Solitaire gameplay variables and player demographic information. To fit each of these models, we used outcomes drawn from either the validation work done from each of the included scales, or from the documentation sent directly from TMB. We looked to three main sources in the development of our model:

- "The TestMyBrain Digital Neuropsychology Toolkit: Development and Psychometric Characteristics" (Singh et al., 2021);
- "Core Neuropsychological Measures for Obesity and Diabetes Trials: Initial Report" (D'Ardenne et al., 2020);
- Documentation provided directly from TMB.

Because the models were fit to data with players grouped within GoTD Solitaire hands, it was likely that there was some degree of statistical dependency in the data that needed to be accounted for in order to avoid biasing our statistical inferences. To account for these statistical dependencies, we estimated cluster-robust variances in *R* using the *sandwich* package (Zeileis et al., 2020). The standard errors were defined as follows:

$$Var(\epsilon_i \mid Hand_i) \; = \; \sigma^2 \, \rho_{Hand} \tag{2}$$

where $\boldsymbol{Var(\epsilon_i \mid Hand_i)}$ represents the variance of the error term (ϵ_i) conditional on the Solitaire hand received by player *i*, $\boldsymbol{\sigma^2}$ is the common within-group variance, and $\boldsymbol{\rho_{Hand}}$ is the intragroup correlation coefficient, representing the correlation of errors within the same group. Once each of these models was fit, they were evaluated for statistical and practical significance.

Each regression model fit to each TMB scale was assessed for its predictive power via R^2, i.e., the coefficient of determination. Using R^2, we selected the models that had the highest R^2 and thereby determined which facets of mental acuity were more accurately predicted by Solitaire gameplay. In addition, we also evaluated the regression coefficients for each model for statistical significance to help us determine which variables are important to predicting mental acuity across scales.

Checking Model Assumptions

As we explored each model, we also considered the assumptions of linear regression:

- Linearity: The relationship between the independent and dependent variables is linear. This was checked by looking at a plot of the residuals—residuals should be randomly scattered around 0 if linearity holds.
- Independence: Observations are independent of each other. This cannot be checked with a test, but we know for a fact that players playing the same game will not be fully independent. With this in mind, we used cluster-robust variances in order to account for the dependencies in those outcomes.
- Homoscedasticity: The variance of errors is constant across all levels of the independent variables. This was checked using a plot of the model residuals. Being scattered equally around the line suggests that this assumption holds.
- No multicollinearity: Independent variables are not highly correlated with each other. To check this assumption, we calculated the variance inflation factor (VIFs) and analyzed it to determine if any variables had problematic collinearity.
- Normality of errors: The residuals are normally distributed. To check this assumption, we visually analyzed a Q-Q plot. Residuals should follow a straight line if normality holds.

If the model assumptions of homoscedasticity, linearity, or normality of errors did not hold, transformations were applied to the data. If there was obvious multicollinearity, the variance inflation VIF was calculated for each independent variable, and problematic variables were either combined or removed (when appropriate).

Validation Approach

To validate our methodology, we examined for convergent validity by correlating predicted cognitive performance scores from our regression model(s) with players' self-reported cognitive impairment, expecting statistically significant negative relationships.

Table 5

Convergent Validity Analyses

Validation question	Analyses	Expected results
What is the overall relation between players' actual mental acuity scores and their predicted mental acuity scores from our model?	The data are split into a training sample and cross-validation sample. The model parameters are estimated using the training sample. The estimated parameters are then applied to the cross-validation sample to calculate predicted acuity scores. The predicted acuity scores and the observed acuity scores are then compared.	Players' predicted mental acuity scores will be statistically significantly and highly correlated with players' actual mental acuity scores (from the TMB scales).
What is the overall relation between self-reported cognitive impairment and overall mental acuity score?	Test of group differences (self-reported with and without self-reported cognitive impairment) on overall mental acuity scores.	Players reporting cognitive impairment should have significantly lower mental acuity scores than players not reporting cognitive impairment.
What is the relation between self-reported age and overall mental acuity score?	Correlate self-reported age and mental acuity scores.	There should be a significant negative relation between age and mental acuity scores.
How does the pattern of results compare to the pattern of results reported in Gielis, Vanden Abeele, Verbert, et al. (2021)?	Inspect the magnitude and direction of differences between players who self-reported cognitive impairment vs. those not impaired, and also by age.	The gameplay variables should be consistent with the results reported in Gielis, Vanden Abeele, Verbert, et al. (2021).

Results

Outlier Analysis

After data collection was complete, there were a total of 4,155 collected responses, ranging from 632 to 1,117 players per TMB test. However, upon initial inspection of the dataset, it was clear that two variables were particularly problematic: completion time for Solitaire games and completion time for the TMB tests. In both cases, several players took markedly longer to complete the "Ultra Brief" tests or the Solitaire games, with completion times ranging in the hours. For instance, one Solitaired player took 24 hours to complete the game of the day, suggesting they opened the game, played it for some period of time, and then came back later to finish it.

High completion times pose two major problems to our analysis:

- They likely do not reflect the effortful play patterns we would expect from players looking to generate a mental acuity score from their Solitaired gameplay;
- A large gap in time before finishing the Solitaire game and finishing the TMB test suggests that players may not have been in the same mental state when completing both activities, which in turn could limit the predictive power of the regression models.

Outliers were identified as being greater than 1.5 interquartile ranges below and above the first and third quartiles, respectively (i.e., Q1 - 1.5 × [Q3-Q1] and Q3 + 1.5 × [Q3-Q1]). Applying these criteria to each player's TMB test and their Solitaire gameplay time, 508 players were excluded based on their completion time on the TMB test, and 283 players were excluded based on their game completion time. The final sample size is 3,647, which is 88% of the original sample. See Appendix J for a detailed description of the outlier analysis.

Sample Summary

Table 6 displays a summary of the final sample based on the independent variables relevant to the final model described above. For the most part, the samples are fairly similar for each test.

Table 6

Summary of Final Sample for Each TMB Test

Test name	*n*	Median game completion time (s)	Median hint count	Media n undo count	Median game difficulty	Median age	Proportion male
Choice reaction time	555	197	153	0	74	54.0	.39
Simple reaction time	840	197	65	0	70	60.5	.40
Digit symbol matching	707	196	93	0	70	55.0	.37
Flicker change detection	568	199	166	0	70	62.0	.41
Matrix reasoning	977	201	193	0	0	70.0	.39

Below is a correlation table for all the independent variables from Solitaired (Table 7). The only variables with particularly strong correlations are game completion time and mean time per move. This may pose an issue later, as collinearity can obfuscate the statistical and practical significance of otherwise important variables. We will address this with more scrutiny as we run the regression models.

Table 7

Pearson Correlations Among Quantitative Independent Variables

Variable	Game completion time	Mean time per move	Hint count	Undo count	Game difficulty
Mean time per move	.95	–	–	–	–
Hint count	.07	.04	–	–	–
Undo count	.09	-.07	.06	–	–
Game difficulty	-.01	.03	-.04	-.15	–
Age	.27	.34	-.08	-.08	.07

Variable Transformations

As noted in the Method section, to improve the linear regression, it was helpful to transform both the Solitaired and TMB variables to be closer to normal. Although linear regression does not have any assumptions regarding the normality of the independent or dependent variables, it does have assumptions regarding the normality of the residuals. Normalizing the independent variables is one way to potentially improve the model relative to the assumptions of linear regression. Table 8 summarizes the transformations performed on TMB variables and Solitaired game variables and Table 9 presents correlations between the Solitaired gameplay variables and the TMB variables. Appendix K contains a detailed analysis of the variables.

Table 8

Transformed Variables and Type of Transformation Used

Variable	Transformation procedure	Used in final model?
TMB variables		
Choice reaction time (mean reaction time)	Log transformation	Yes
Simple reaction time (mean reaction time)	Log transformation	No
Digit symbol matching (number correct)	None	Yes
Flicker change detection (score)	Log transformation	Yes
Matrix reasoning	None	No
Solitaired gameplay variables		
Game completion time	Log transformation	No
Mean time per move	Log transformation	Yes
Hint count	Dichotomized	Yes
Undo count	Dichotomized	No
Game difficulty	Squared transformation	No

Table 9

Correlations Between Solitaired Gameplay Variables and TMB Variables Used in Final Model

Solitaired gameplay variables	Choice reaction time	Digit symbol matching	Flicker change detection
Log of mean time per move	.37**	-.38**	-.41**
Hints (binary)	-.04	.05	.03
Undos (binary)	-.08	.11**	-.01
Game difficulty	.05	-.12**	-.02
Age	.72**	-.72**	-.67**
Gender: Male	-.03	-.05	.07

**p* < .05. ***p* < .01.

Regression Modeling

We developed the regression in two stages. The first stage was to examine regression models for each mental acuity measure (i.e., Choice Reaction Time, Digit Symbol Matching, and Flicker Change Detection) with the same set of independent variables for all models (i.e., *game completion time*, *mean time per move*, *hint count*, *undo count*, *game difficulty*, *age*, and *gender* [male]). The three models represent a different aspect of mental acuity and together they would offer a fuller picture of a player's capabilities. Because each mental acuity measure had from one to three measures (e.g., Choice Reaction Time had mean reaction time, median reaction time, and accuracy of response), we examined each measure with respect to model assumptions (i.e., linearity, multicollinearity, normal distribution of residuals, homoscedasticity) and model fit indices.

In all three models, both *age* and *mean time per move* are significant. In two models, *gender* is a significant variable; however, it is only marginally significant in both. One additional advantage of the Flicker Change Detection score is that it is the only model in which the use of *hints* is significant, suggesting it more holistically uses the data produced by Solitaired to produce the acuity score. All three models satisfied the model assumptions of regression, suggesting that all of these models are viable.

The second stage was to develop more parsimonious models. In almost all the previous models, neither win percentage nor undo use were significant predictors of mental acuity. Additionally, hint use was only significant in one of the models. We fit follow-up models for Choice Reaction Time, Digit Symbol Matching, and Flicker Change Detection. In the interest of keeping this metric open to all players (not just those in the gender binary), we also removed the gender variable, despite its marginal significance. A hierarchical regression analysis was conducted for the original model and a reduced model. In each case, the removal of various variables (i.e., *mean time per move*, *hint count* [retained for Flicker Change Detection], *undo count*, *game difficulty*, and *gender*) resulted in a less than 1 percentage point drop in the variance explained by the model. This more parsimonious model removes the need for Solitaired to collect player gender information and solely asks players to submit their age before receiving their acuity score. The model fit statistics and indices are show in Table 10.

Table 10

Regression Model Fit Indices

	Choice reaction time	Digit symbol matching	Flicker change detection
N	555	707	568
R^2	.525	.54	.479
Adjusted R^2	.523	.539	.476
Residual Std. Error	0.281 (*df* = 552)	4.915 (*df* = 704)	0.318 (*df* = 564)
F Statistic	305.281** (*df* = 2; 552)	413.742** (*df* = 2; 704)	172.688** (*df* = 3; 564)

*p < .05. **p < .01.

Model Equations[3]

The final model equations are given in Equations 3, 4, and 5. The model inputs are (a) the player's age; (b) whether the player used any hints; (c) the total moves the player took; and (d) the amount of time it took the player to complete the hand. Entering these inputs into the model will generate predicted acuity scores that could be reported in multiple ways (e.g., as a panel of scaled scores, percentiles, or something else entirely).

Equation for predicting Choice Reaction Time:

$$\log\left(\widehat{Mean\ RT}\right) = 6.40 + \text{[redacted]} * \log(Mean\ Time\ per\ Move) + \text{[redacted]} * Age \tag{3}$$

Equation for predicting Digit Symbol Matching Reaction Time:

$$\#\ \widehat{Correct} = 27.09 - \text{[redacted]} * \log(Mean\ Time\ per\ Move) - \text{[redacted]} * Age \tag{4}$$

Equation for predicting Flicker Change Detection score:

$$\log\left(\widehat{Score}\right) = 2.73 - \text{[redacted]} * \log(Mean\ Time\ per\ Move) - \text{[redacted]} * (Hint\ Use) - \text{[redacted]} * Age \tag{5}$$

Validity Checks

In this section, the following validity checks are done visually. First, for each model, the predicted score is compared to the actual score. Second, for each model, two distributions are compared: the predicted output of the model using the training data and the predicted outputs

[3] For more information on the model coefficients, please contact the authors.

of the model using the hold-out data (i.e., data not used to train the model). Third, for each model, the two distributions (i.e., using training data and hold-out data) are then plotted as a function of their input variables.

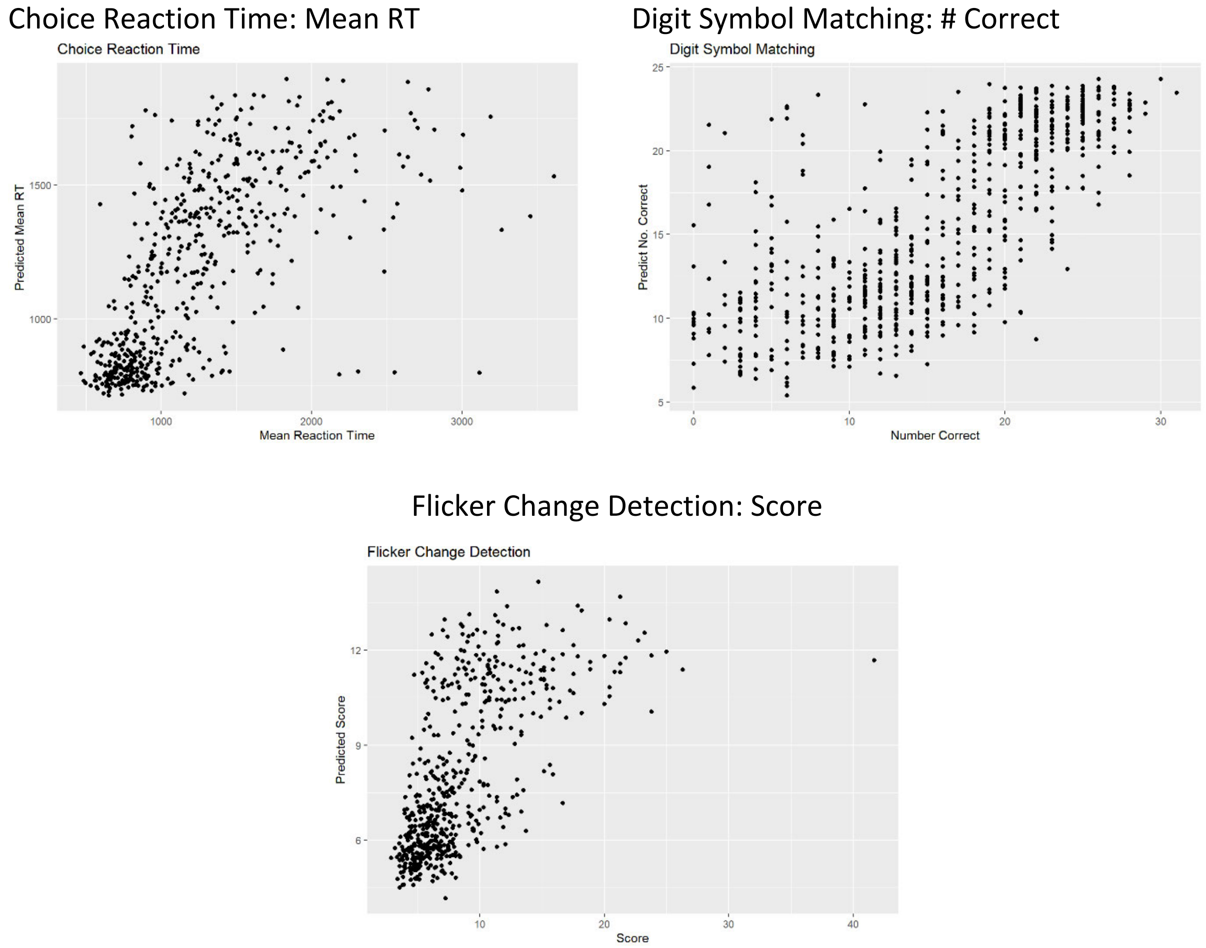
The next set of validity checks examine the predicted output and players' self-reported cognitive impairment. We examined player's self-reported cognitive impairment and determined whether its relationship with each variable is consistent with (a) the pattern of significance in each of the final models and (b) the noted variables from Gielis, Vanden Abeele, Croon, et al. (2021) and Gielis, Vanden Abeele, Verbert, et al. (2021).

Comparison of Actual and Predicted TMB Scores

Figure 1 depicts the relationship between the actual and predicted outcomes. As we can see, in all three instances there is a fairly strong relationship with the exception of a couple of notable players.

Figure 1

Actual vs. Predicted Outcomes From Final Model



Choice Reaction Time: Mean RT

Digit Symbol Matching: # Correct

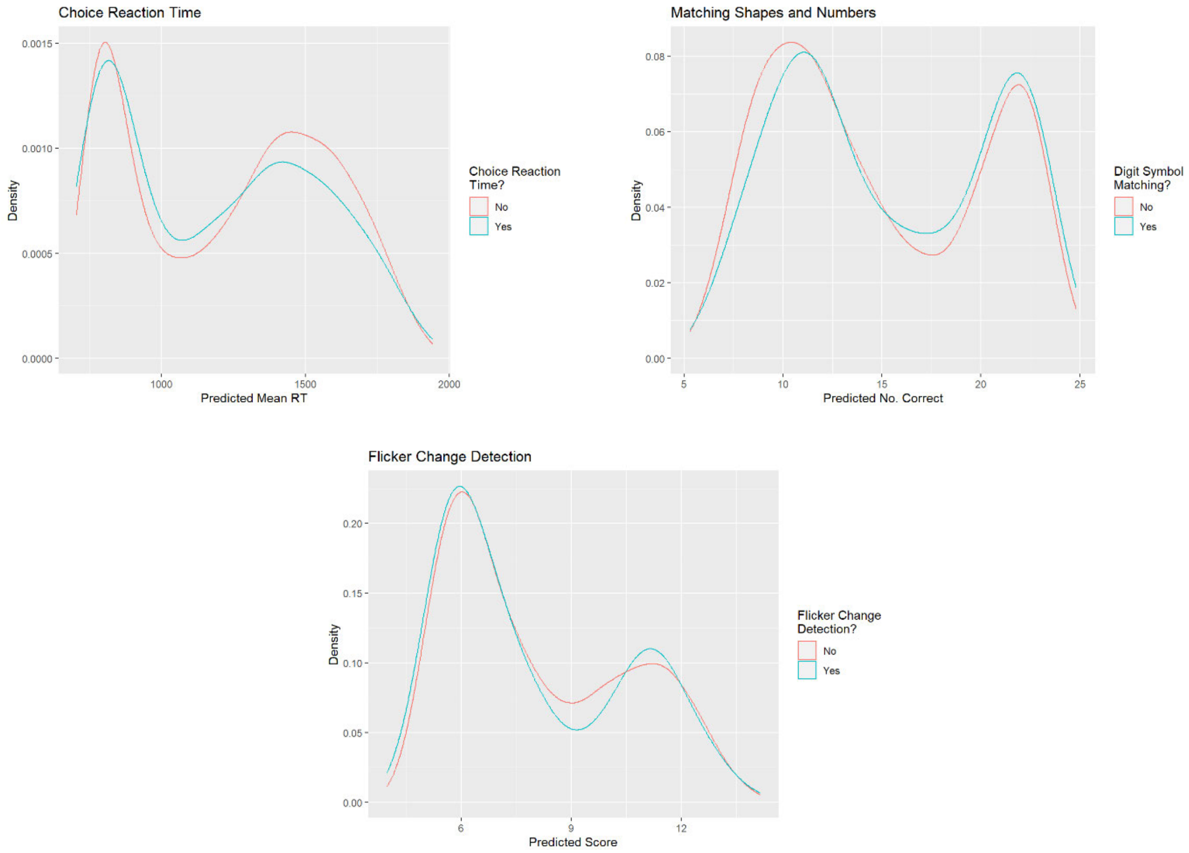
Flicker Change Detection: Score

It is unclear why one player scored so poorly in the predicted Flicker Change Detection model as compared to their TMB scores; however, it is likely that because so few players scored so highly on Flicker Change Detection, the model did a poor job accounting for them. Fitting a model without this individual, though, sees a less than 1% increase in R^2. Consequently, whether or not this individual is included in the final model is a judgment call but we recommend including it. Outliers represent a potential limitation of the model: Because they were underrepresented in our model, there may be some potential range issues. However, the model performs fairly well for all players, so these edge cases are likely worth ignoring.

Visualizing the Model With the Full Dataset. In this section, we explore the relationship between the predicted outcomes to the individual variables for each model using the full dataset. Figure 2 depicts the distribution of the predicted outcomes using the whole dataset. For each model, the distribution for the training data (which was used to fit the model) and the rest of the dataset (which was not used to fit the model) are overlaid. In all three, one can observe that the training data and the remaining responses closely match in terms of predicted outcome. Notably, each of these models closely mirrors the U-shaped distribution of player age (Figure K6). This would make sense, as age is an important factor in predicting mental acuity.

Figure 2

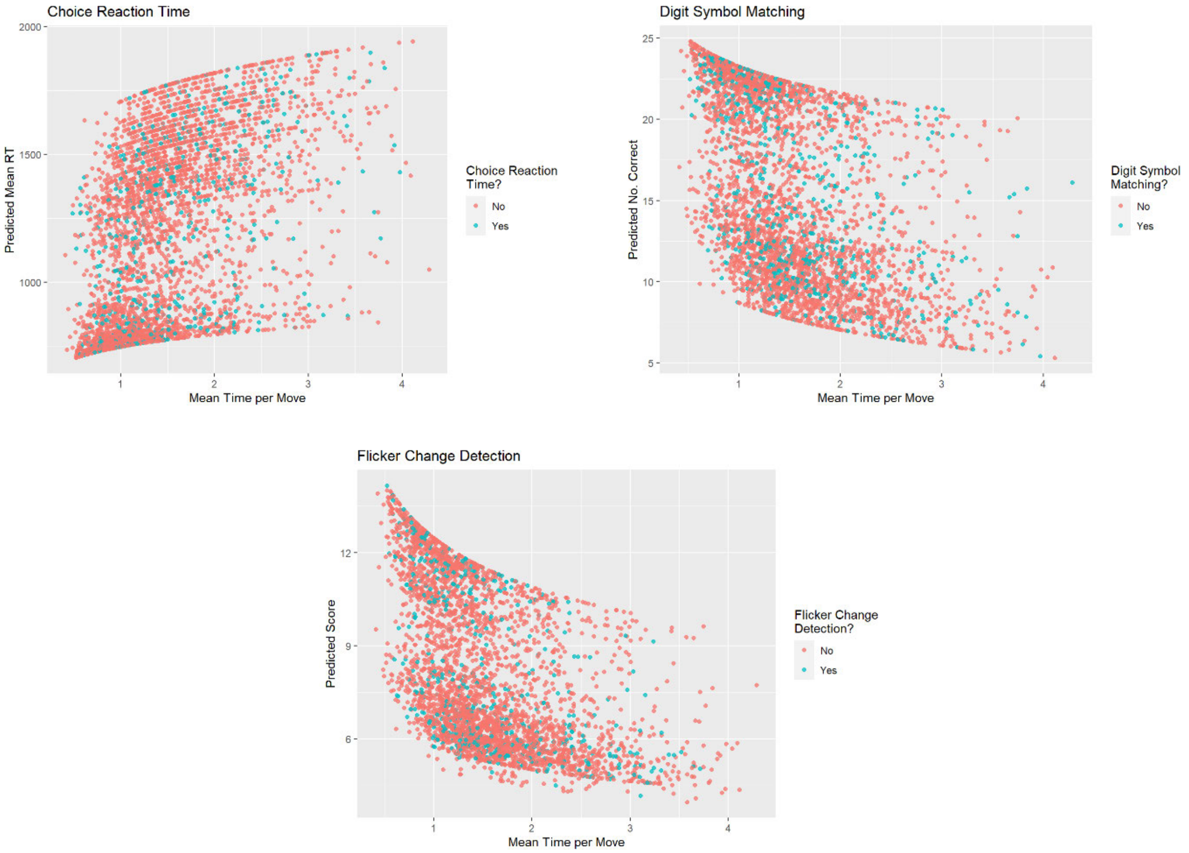
Predicted Mental Acuity Outcomes for Full Dataset



Predicted Acuity vs. Mean Time per Move. Generally speaking, as *mean time per move* increases, the predicted Choice Reaction Time *mean reaction time* increases and the Digit Symbol Matching *score* and Flicker Change Detections *score* decrease. This relation does not appear to be linear, which is expected, given that the model was fit on a transformed time variable. Once again, we can see that the distributions mirror each other (Figure 3).

Figure 3

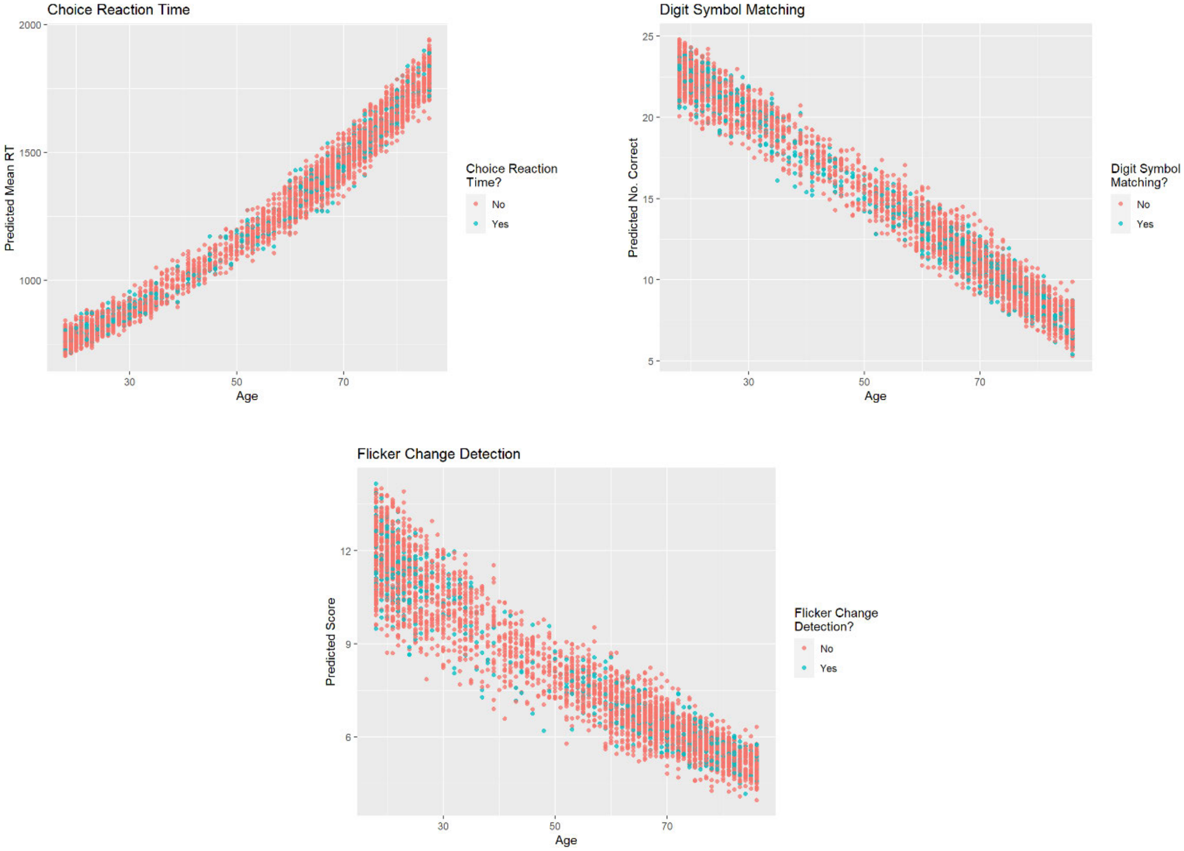
Predicted Mental Acuity Outcomes for Full Dataset vs. Mean Time per Move



Predicted Acuity vs. Age. There is no discernible pattern difference here between the data used for modeling and the rest of the dataset (Figure 4).

Figure 4

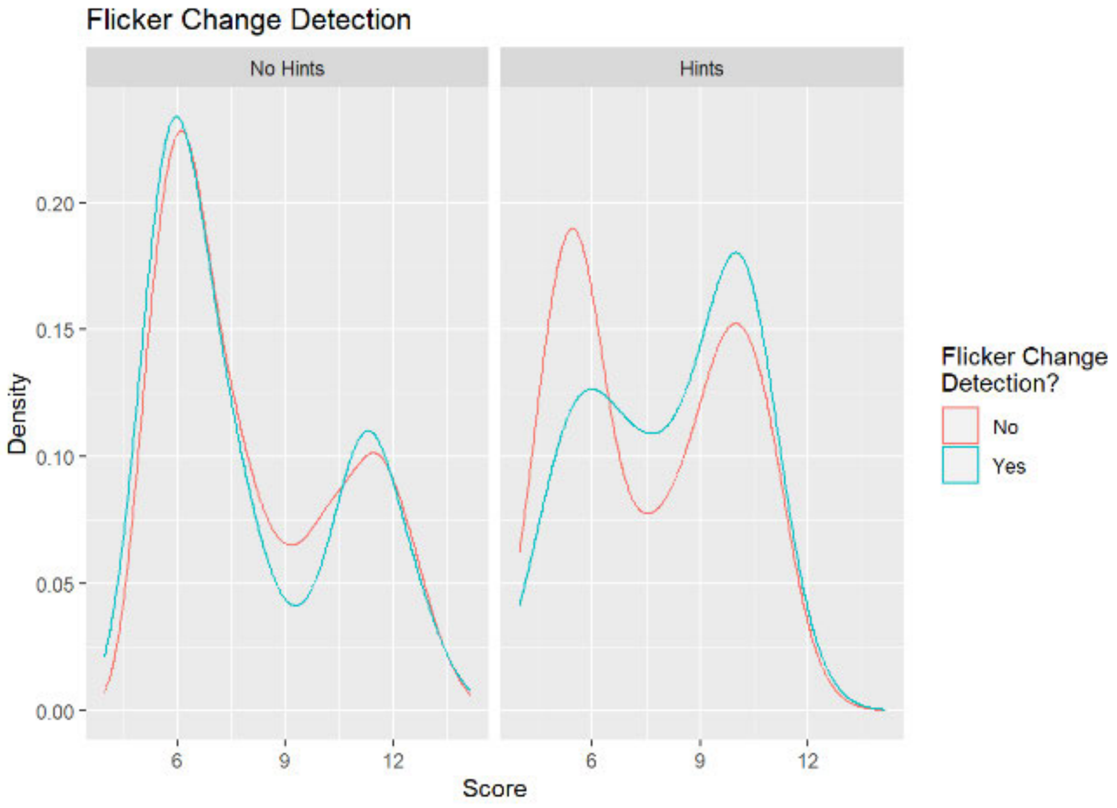
Predicted Mental Acuity Outcomes for Full Dataset vs. Age



Predicted Acuity vs. Hint (Binary). Figure 5 depicts the distribution of predicted acuity scores for the players who used hints and those who did not. This analysis was only done for the Flicker Change Detection model, as it was the only model in which hint use was significant.

Figure 5

Predicted Mental Acuity Scores for Full Dataset Grouped by Hint Use



This plot is somewhat challenging to interpret, but essentially, the plot on the left is the distribution of predicted scores for those who did not use hints, and the plot on the right is the distribution of predicted scores for those who used hints. The colored lines represent whether the data came from the training data or the rest of the dataset. Here we can see that there did appear to be a small discrepancy between the training and nontraining data for those who used hints. However, the general trend (a bimodal distribution with peaks around 6 and 10) appeared in both, and the proximity of the means for the training and nontraining data (8.27 and 8.57, respectively) suggests this is not a large concern.

Comparison of Predicted TMB Scores With Self-Reported Cognitive Impairment

For the validity checks of our model, we turn to player self-reported cognitive impairment (Table 11). In this section, we look at player self-reported cognitive impairment and determine whether its relationship with each variable is consistent with (a) the pattern of significance in each of the final models and (b) the noted variables from Gielis, Vanden Abeele, Croon, et al. (2021) and Gielis, Vanden Abeele, Verbert, et al. (2021). As a note, many players may not feel comfortable disclosing cognitive impairment, so it was not a contender as a covariate in our modeling.

Table 11

Number of Players Self-Reporting Cognitive Impairment by TMB Test

TMB test name	Impairment?	Frequency
Choice reaction time	No	495
	Yes	60
Digit symbol matching	No	638
	Yes	69
Flicker change detection	No	503
	Yes	65

Relationship Between Actual Scores to Self-Reported Cognitive Impairment. Table 12 depicts a model in which self-reported cognitive impairment is used as the sole variable to predict Choice Reaction Time, Digit Symbol Matching, and Flicker Change Detection scores. We would expect a significant relationship with self-reported cognitive impairment. This model suggests that there is a relationship between self-reported cognitive impairment and Flicker Change Detection scores. However, the poor model fit implies it is a dubious relationship, especially among the other two nonsignificant models.

Table 12

Model Predicting Relationship Between Self-Reported Cognitive Impairment and Final Model Outcomes

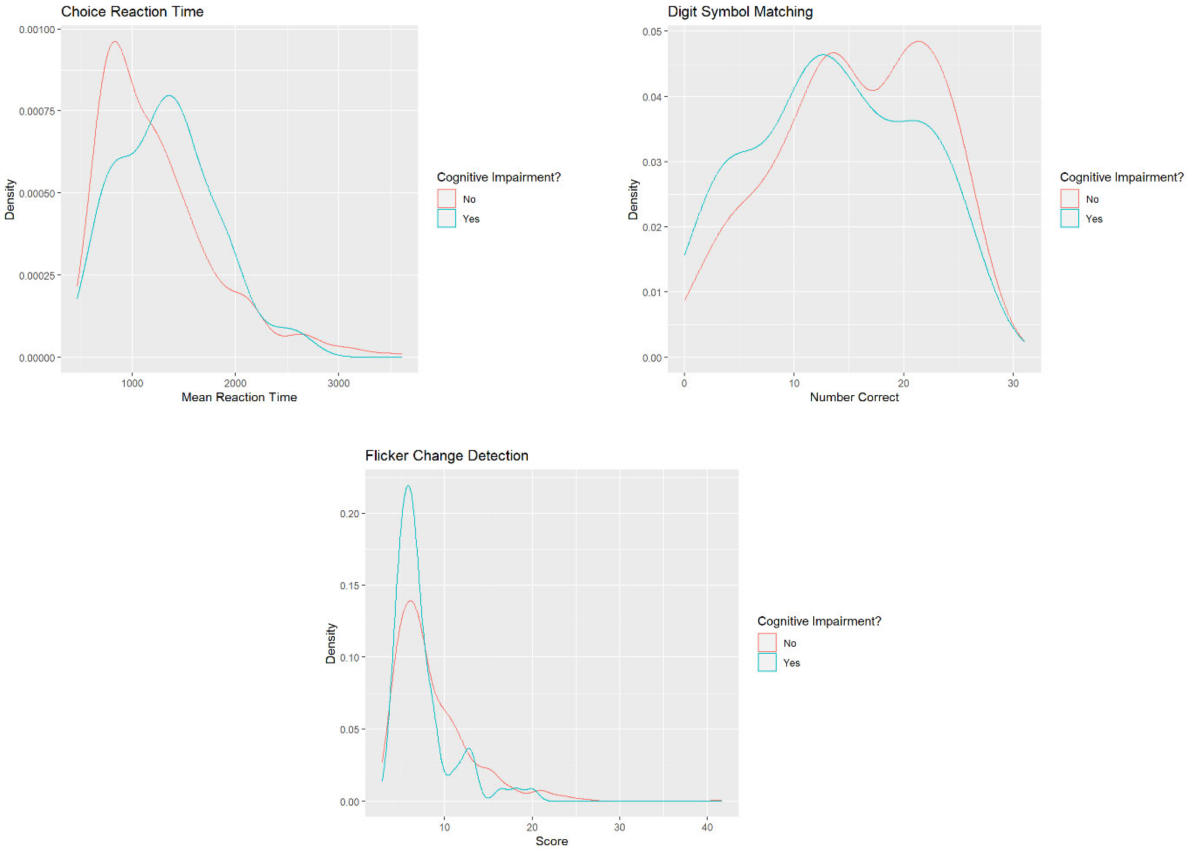
	Dependent variable		
	Choice reaction time	Digit symbol matching	Flicker change detection
Constant	1,236.270**	15.600**	2.044**
	(-30.139)	(-0.328)	(-0.021)
Cognitive impairment	104.20	-1.934*	-0.123*
	(68.64)	(-0.965)	(-0.05)
Observations	555	707	568
R^2	0.004	0.006	0.008
Adjusted R^2	0.002	0.005	0.006
Residual Std. Error	546.731 (*df* = 553)	7.221 (*df* = 705)	0.438 (*df* = 566)
F Statistic	1.944 (*df* = 1; 553)	4.465* (*df* = 1; 705)	4.565* (*df* = 1; 566)

*p < .05. **p < .01.

Looking to Figure 6, we see that there is relatively little visual difference between the distributions for those reporting cognitive impairment and those not. Turning to Wilcoxon tests, we see a very weak relationship. For Choice Reaction Time, the difference in means for the group reporting cognitive impairment (*M* = 1340.47) and those not (*M* = 1236.27) is marginally significant [*W* = 12262, *p* value = .027], meaning there is some evidence that those who reported cognitive impairment reacted slightly more slowly. For Digit Symbol Matching, the difference reporting cognitive impairment (*M* = 13.67) and those not (*M* = 15.6) is also statistically significant [*W* = 25242, *p* value = .045]. Lastly, for Flicker Change Detection, the difference between those reporting cognitive impairment (*M* = 7.36) and those not (*M* = 8.59) is statistically significant [*W* = 19169, *p* value = .023]. As a note, we use a Wilcoxon test here as it is robust to the nonnormality of the data.

Figure 6

Distribution of Mental Acuity Scores Grouped by Self-Reported Cognitive Impairment



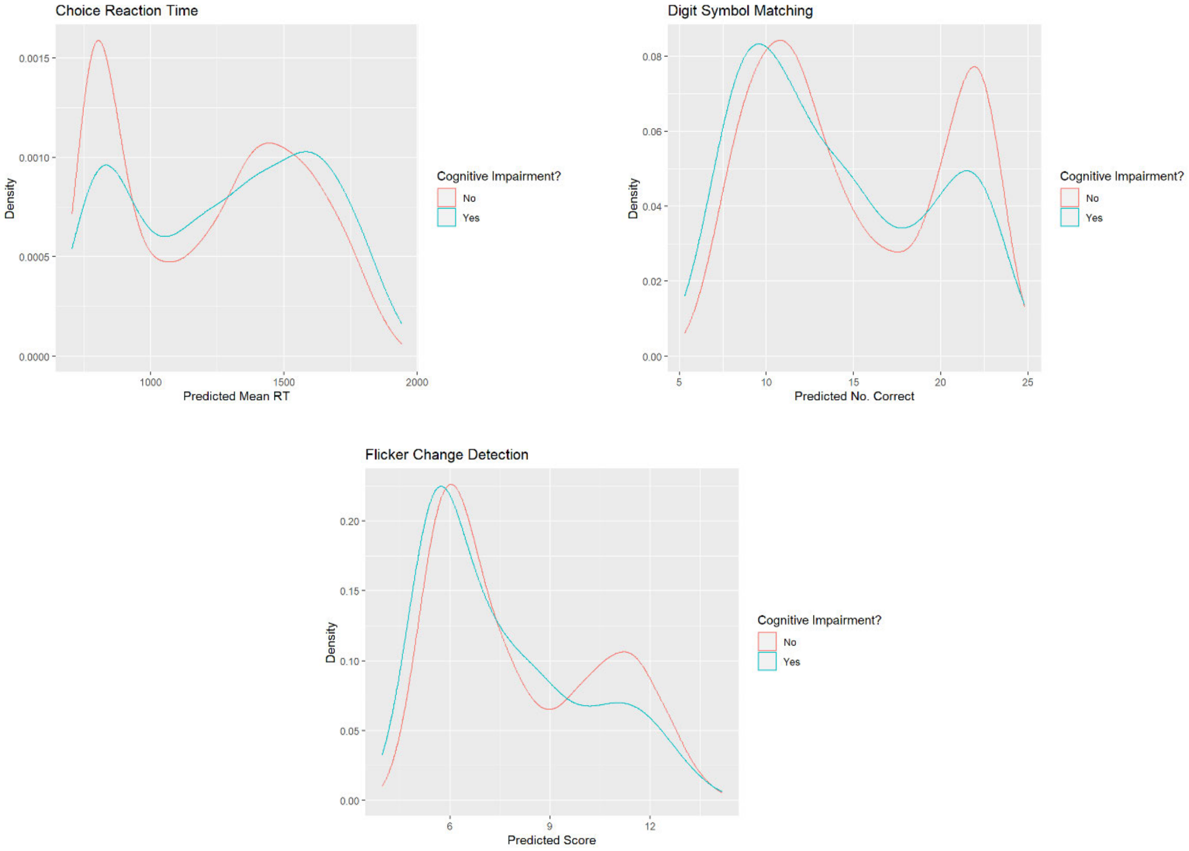
Relationship Between Predicted Outcomes to Self-Reported Cognitive Impairment. Figure 7 depicts the entire sample, including those who did not take the TMB tests in question. Based on these plots and Wilcoxon tests, we can see that the outcomes follow the pattern we would both expect and hope for:

- Predicted Choice Reaction Time response time is significantly higher for those reporting cognitive impairment (*M* = 1287.22) compared to those not (*M* = 1223.89) [*W* = 551894, *p* < .001],
- Predicted Digit Symbol Matching score is significantly lower for those reporting cognitive impairment (*M* = 14.02) compared to those not (*M* = 15.03) [*W* = 689249, *p* < .001],
- Predicted Flicker Change Detection score is significantly lower for those reporting cognitive impairment (*M* = 7.63) compared to those not (*M* = 8.13) [*W* = 696872, *p* < .001].

In all three cases, the difference is small, but the significance of it implies that the model itself can differentiate between people with and without self-reported cognitive impairment to some degree.

Figure 7

Distribution of Predicted Mental Acuity Grouped by Self-Reported Cognitive Impairment



Relationship Between Self-Reported Cognitive Impairment and Other Variables Considered for the Final Model. Table 13 depicts the relationship between self-reported cognitive impairment and the other variables in the final model, looking at the full dataset. In order to provide some validity evidence for our model, we would hope that the following statistical tests follow a similar pattern of significance to the final model.

Table 13

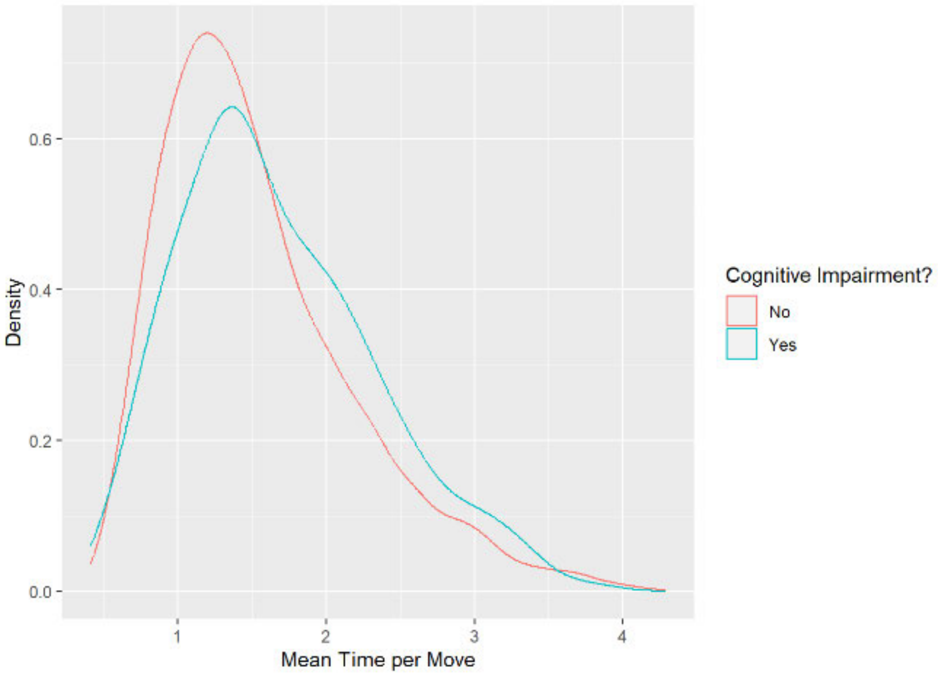
Summary of Potential Model Variables Grouped by Self-Reported Cognitive Impairment

Self-report?	*n*	Mean time per move (*SD*)	Proportion of players who used hints	Proportion of players who used undos	Mean game difficulty (*SD*)	Mean age (*SD*)	Proportion female
No	3268	1.54 (0.65)	.10	.37	66.93 (18.35)	52.49 (23.06)	.60
Yes	379	1.67 (0.66)	.13	.34	67.9 (17.51)	56.25 (22.2)	.64

Mean Time per Move. We would expect this relationship to be significant based on the final model and Gielis, Vanden Abeele, Verbert, et al. (2021). Based on visual inspection of Figure 8 and a Wilcoxon rank test, *mean time per move* is significantly higher for people self-reporting cognitive impairment [W = 538801, p < .001].

Figure 8

Distribution of Mean Time per Move Grouped by Self-Reported Cognitive Impairment



Hint Use (Binary). Because we are looking at whether players used hints at all, this cannot be presented visually as a graph. Instead, we present it as a proportion table (Table 14) along with an associated proportion test. It is unclear from our models whether this should be significant, as it is not significant in the Choice Reaction Time or Digit Symbol Matching models, but it is in the Flicker Change Detection model. Based on the results of a proportion test, we do not find significance [χ^2 = 1.80, *df* = 1, *p* value = .18]. This is in line with Gielis, Vanden Abeele, Croon, et al. (2021), in which hint use was found to be nonsignificant.

Table 14

Proportion of Players Self-Reporting Cognitive Impairment Using Hints

Self-report?	Did not use hints	Used hints
No	.90	.10
Yes	.87	.13

Undo Use (Binary). We would expect this to not be significant both based on the final models and on Gielis, Vanden Abeele, Croon, et al. (2021). We once again use a proportion table (Table 15) and proportion test, and find no significant difference in proportions [χ^2 = 0.81, *df* = 1, *p* value = .37].

Table 15

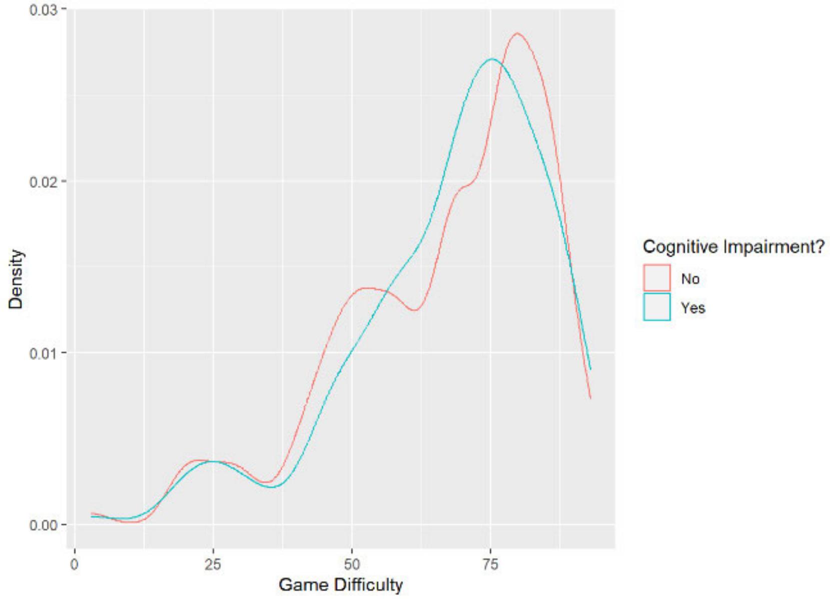
Proportion of Players Self-Reporting Cognitive Impairment Using Undo

Self-report?	Did not use undos	Used undos
No	.63	.37
Yes	.66	.34

Game Difficulty. The distribution of game difficulty is depicted in Figure 9. This variable was not studied by Gielis and it was not studied in the pilot. Although theoretically we would expect to see this to be significant based on expert judgment, it is not significant in the final models. Consequently, we would expect to see no relationship between these variables, which appears to be the case based on a Wilcoxon Rank Test [*W* = 607439, *p* = 0.54].

Figure 9

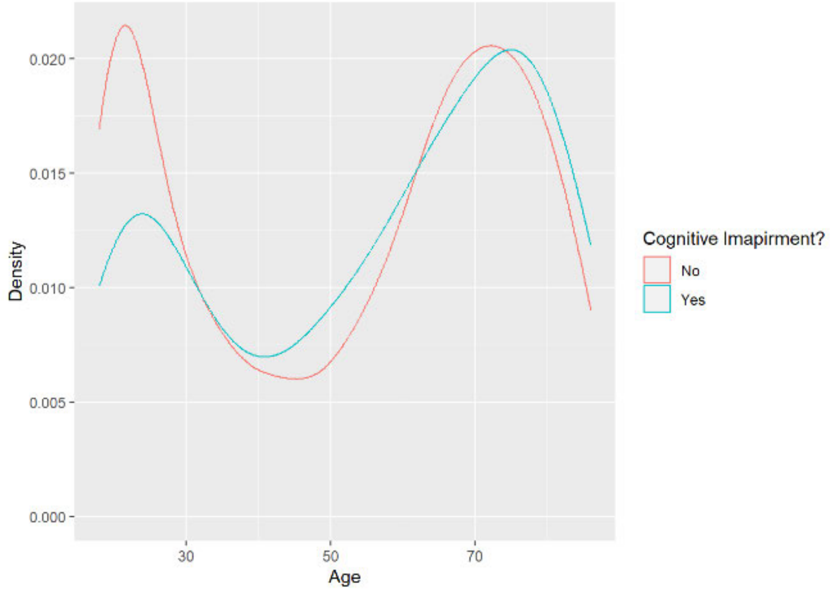
Distribution of Game Difficulty Grouped by Self-Reported Cognitive Impairment



Age. As the most consistently significant variable in all of our models, we would hope to see a significant relationship here. Figure 10 depicts this relationship visually, and the Wilcoxon test confirms our expectations [*W* = 559251, *p* = .002].

Figure 10

Distribution of Age Grouped by Self-Reported Cognitive Impairment



Gender. It is unclear what pattern we should look for here, as this variable was significant in two of the final model candidates and not in the other. We opted to not include it in the final models for practical reasons. However, looking at the proportion table (Table 16) and the results of the proportion test below, we see that this is not the case [χ^2 = 1.82, *df* = 1, *p* value = .178]. This provides additional validation for its exclusion in the final models.

Table 16

Proportion of Players Self-Reporting Cognitive Impairment by Gender

Self-report?	Female	Male
No	.60	.40
Yes	.64	.36

Discussion, Limitations, and Caveats

This study was based on the existing scientific literature on the use of Solitaire gameplay measures to differentiate between a healthy group of individuals and a group of individuals diagnosed with mild cognitive impairment. Our pilot study involved existing Solitaired gameplay measures and Solitaired players showed results consistent with the existing literature. These two factors increased our confidence in the idea of using Solitaire gameplay as a way to report on players' mental acuity.

The results of the study indicate a strong and significant relationship between information collected by Solitaired.com (i.e., *mean time per move* gameplay variable, self-reported age) and aspects of mental acuity, such as:

- Processing speed, response selection/inhibition, and attention (measured by the Choice Reaction Time TMB test),
- Processing speed and visual short-term memory (measured by the Digit Symbol Matching TMB test).
- Visual search, change detection, and visual working memory (measured by the Flicker Change Detection TMB test).

The validity checks of the final models support these relationships, as the significant variables in the final model align with the significant and nonsignificant variables found in the research of Gielis and colleagues (Gielis, Vanden Abeele, Croon, et al., 2021; Gielis, Vanden Abeele, Verbert, et al., 2021). In all the proposed models, *mean time per move* was highly significant and of moderate magnitude. On the other hand, as expected, variables such as *undo* use and *hint* use were not found to be significant (with the exception of the Flicker Change Detection model). However, not all validity checks supported the final models: self-reported cognitive impairment was only weakly related to the predicted model outcomes (Table 13). This weak relationship, though, may be due in part to the poor quality of the self-reported cognitive impairment measure. It is possible that players were apprehensive to report this sensitive information, which is supported by its poor and nonsignificant relationships to TMB outcomes (Table 12).

Before moving on to the discussion of how scores are reported to players, it is prudent to examine whether there is an additional benefit to providing the players with scores generated from three distinct models or whether generating an acuity score from fewer models provides the same statistical information. This consideration is particularly poignant if player scores are presented normatively, such as percentiles. To determine if more than one model is necessary, players' scores from each of the final three models were correlated (Table 17).

Table 17

Pearson Correlations of Final Model Scores

Model	Choice reaction time	Digit symbol matching
Digit symbol matching	1.00	–
Flicker change detection	.98	.99

While the correlations among models were extremely high, suggesting that using only one model would suffice for reporting, high correlations do not confirm that the models provide equivalent information. Correlations do not account for differences in scale or potential nonlinear distortions between the score distributions. Thus, we checked to see how differently the three models ranked players. If the three models produced similar rankings, then one model would suffice for reporting purposes.

To check the rankings of the three models, scores from each model were converted into percentile ranks, ensuring comparability across scales, and the absolute difference in percentile ranks for each player was calculated. The results showed that, on average, players' ranks differed by only about 1-3 percentile points across models, suggesting minimal practical differences in the ordering of individuals. This supports the claim that the models produce nearly interchangeable information, reinforcing the idea that maintaining all three models may be redundant. Consequently, we recommend that scores are only generated from the final Digit Symbol Matching model, which displays strong model fit indices and the highest R^2 of the final model contenders.

Limitations

In this study, we are focusing on the measurement of aspects of mental acuity through gameplay and not the measurement of mild cognitive impairment (MCI). While the TMB scales represent aspects of cognitive functioning associated with MCI, we are not proposing to develop a game-based measure of MCI. Second, any inferences drawn about mental acuity as measured by the TMB scales are limited to the Solitaired sample. We know very little about how representative the Solitaired sample is to the general population.

Another important long-term consideration is that if the design of the game changes in a way that leads players to respond very differently—for example, changes in the UI/UX, or achievement system—anything that is likely to result in players using different strategies or change the way they interact with the system, then the parameters of the statistical models may need to be updated for the new design. That is, the coefficients of the statistical model will reflect the relation between gameplay behavior and TMB at the time of data collection. If gameplay behavior changes substantially because of changes in the game design, then the coefficients that reflect the relation between gameplay and TMB at the time of data collection may no longer be valid.

Reporting Player Scores

There are a number of potential methods for reporting player scores. The most salient options are presented here. Ultimately, the decision of how to report scores is at the discretion of Unwind Media, however, the recommendation of Lyons Assessment Consulting is that only scores from the final Digit Symbol Matching model are presented to the players and only as percentile scores:

- An overall percentile score (normed on the study data) generated using the final Digit Symbol Matching model presented above.
- A percentile score within player age group.

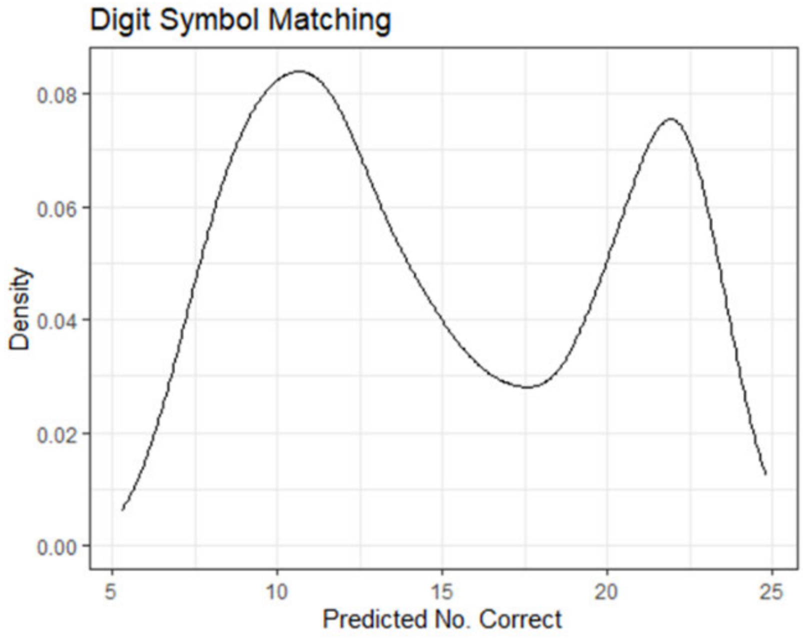
We recommend giving players the option to see their score relative to people in their same age bracket. Because age is such a meaningful variable in the model, older players are, on average, going to have lower mental acuity scores. Consequently, giving those players the option to compare themselves to other players in their age range could be beneficial.

Raw Outcome

If we assume the people who participated in our study are a fair representation of the players of Solitaired.com, then we could expect the following distribution for the final model (Figure 11).

Figure 11

Distributions of Raw Predicted Scores for the Final Model



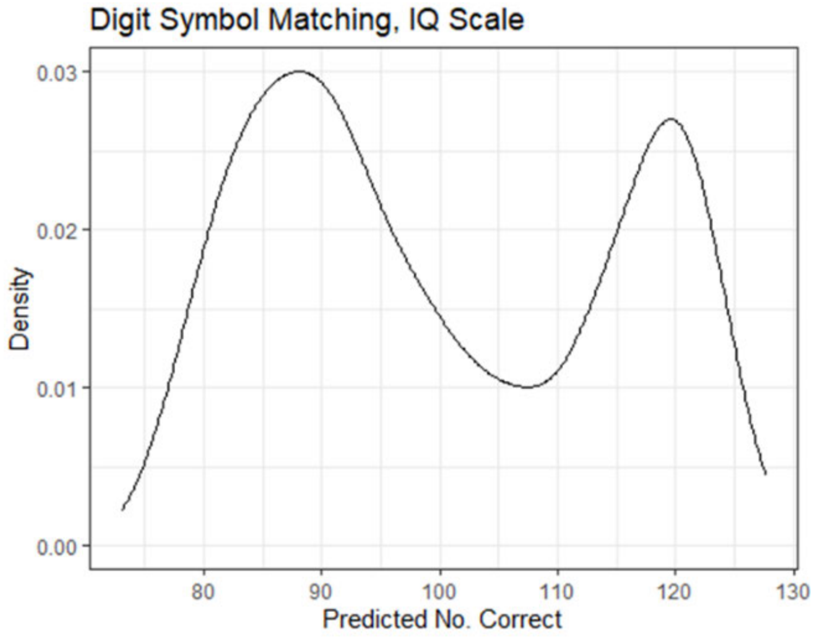
As a note, we flipped the Choice Reaction Time Mean RT scale, as a higher RT is worse than a lower one, and it would make more sense to keep the scales a consistent direction for players. These numbers, however, are devoid of context and present the worst possible option. At best, the interpretation of the numbers would be, "Based on your Solitaired gameplay, we predict you would receive a score of X on a comparable mental acuity test." While this is important to have in our documentation, it is not something that should be presented to players. Instead, the best possible option is to transform the outcome to a score that is more interpretable to players.

Scaled Outcome Using Linear Transformations

There are many possible ways to transform the data to put it onto a scale that is more interpretable to players. We could give players a score on a scale similar to that of an IQ test. The IQ scale is centered at 100 with a standard deviation of 15, as shown in Figure 12.

Figure 12

*Mental Acuity Model Transformed Onto IQ-Style Scale (*M *= 100,* SD *= 15)*



As this is a linear transformation, we can see that the nature of the distribution does not change at all; it is just shifted/compressed to fit the IQ-style scale range. There are other possible scales, such as the SAT ($M \approx 650, SD \approx 210$), a curved test out of 100 ($M = 70, SD = 10$), or some other scaled score. Below we discuss some other options.

Placing players on the new scale is relatively easy given that we have the mean and standard deviation of the norming data that we used to fit the model. If we are using the IQ-style scale, an example of how to convert Digit Symbol Matching scores is presented here. For Digit Symbol Matching, we have a mean of 14.92 and a standard deviation of 5.36, so the conversion model would be as follows:

$$\text{Scaled Score} = \left(\frac{\text{Predicted} - 14.92}{5.36}\right) * SD + M \tag{6}$$

If we were to use an IQ-style scale, we would use *SD* = 15 and *M* = 100.

Normative/Percentile Outcome

One possible outcome, which provides a surprising amount of flexibility, is providing players with a normative score, such as a percentile rank. A percentile is a value on a scale from 0 to 100 that indicates the percentage of data points in a set that are below it. For example, if a player is in the 90th percentile for predicted acuity, that player's predicted score is higher than 90% of the people measured based on Solitaired gameplay. If we treat the initial sample of

3,647 players as our baseline for norming, any predicted score from any new Solitaired game could be associated with a percentile value.

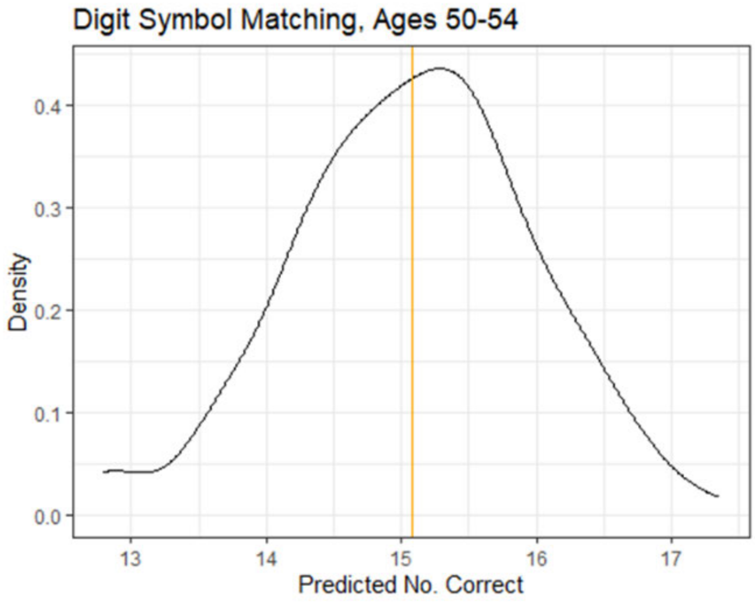
Reporting scores as percentile ranks offers several advantages, particularly when dealing with multiple scores from different scales. First, percentile ranks provide a standardized, intuitive way to interpret scores by expressing a player's standing relative to others, regardless of the underlying measurement scale. This ensures comparability across the three models, allowing players to understand their relative performance without needing to interpret raw scores that may differ in range or distribution. Additionally, percentile ranks mitigate issues related to differences in scale units, making it easier to identify meaningful differences between scores from different models. By placing all three scores on a common, interpretable metric, players can more easily compare outcomes across models.

Relative Normative Scales

Given the strong significance of the *age* variable, it might also be of interest of Solitaired to present players their scores relative to other players in the same age group (our sample was not large enough to have sufficient sample size for each individual age, so age groups are necessary). For instance, looking at players in the 50 to 54 age range, Figure 13 displays the predicted mental acuity scores (the line is drawn for a unique player with an average score in each of the three categories). An interpretation, then, would be, "Relative to other players in your age range, you scored in the top 48% of players for processing speed, the top 51% of players for visual short-term memory, and the top 53% of players for working memory."

Figure 13

Distribution of Predicted Outcomes for Players Aged 50-54



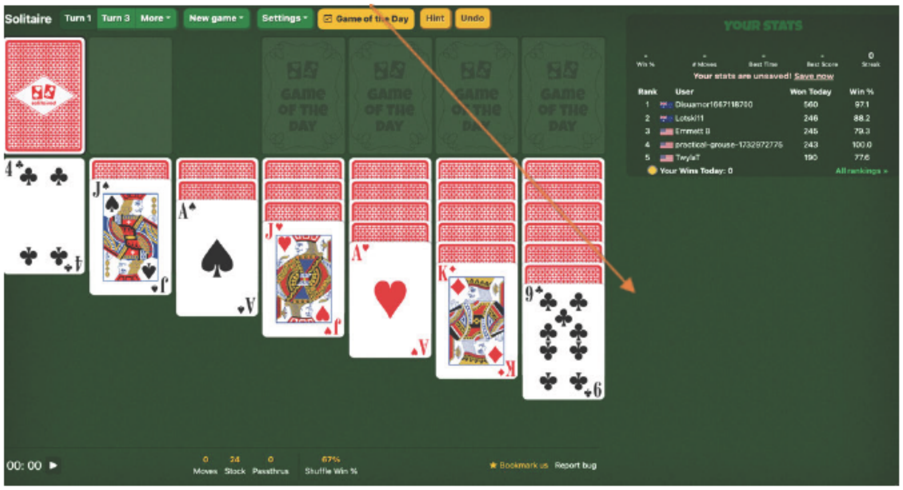
Some Recommendations for Score Reporting

This section draws on previous sections regarding player score reporting to present a set of recommendations for visually incorporating analytic results into Unwind Media's Solitaired web app interface. For the initial presentation to the player, we recommend reporting total cognitive ability percentiles using a horizontal comparative bar chart displayed on the main screen once a game is completed and a player's scores have been calculated.

We recommend this visual persist on the screen for any subsequent games played, positioning it, for example, underneath the "Your Stats" box displayed for the player (see Figure 14).

Figure 14

Possible Location to Report Player Mental Acuity Score



-	-	-	-	0
Win %	# Moves	Best Time	Best Score	Streak

Your stats are unsaved! Save now

Rank	User	Won Today	Win %
1	Dlsuamor1667118760	560	97.1
2	Lotski11	246	88.2
3	Emmett B	245	79.3
4	practical-grouse-1732972776	243	100.0
5	TwylaT	190	77.6

Your Wins Today: 0

All rankings »

0	24	0	67%
Moves	Stock	Passthrus	Shuffle Win %

We also highly recommend Unwind Media include a hyperlink that players can click on to generate a more detailed, fully explained breakdown of player cognitive profiles along with contextual information. This clickable link can generate a separate window that automatically pauses the Solitaire live game for the player and, we believe, should contain the following text:

> **Your Mental Acuity Score**
>
> Based on your gameplay in Solitaired, we've estimated your **Mental Acuity Score** using a statistical model built from a large study of Solitaired players like you. This model was developed by analyzing gameplay patterns alongside psychometrically validated cognitive assessments from The Many Brains Project.
>
> Your score is presented as a **percentile rank**, which tells you how your performance compares to others in our norming sample. We provide two percentile ranks:

> - **Overall Percentile Rank** – Compares your performance to all players in our study.
>
> - **Age-Based Percentile Rank** – A more personalized score that compares your performance to players in your age group.
>
> As an example, if your overall percentile rank is 75, you performed better than 75% of all players in our study. If your age-based percentile rank is 82, you performed better than 82% of players in your age group.
>
> **What This Score Represents**
>
> Your Mental Acuity Score reflects cognitive patterns associated with processing speed, visual short-term memory, and working memory, as observed in Solitaire gameplay. While this score is based on statistical estimates, it provides insight into how your gameplay behaviors relate to broader cognitive abilities.
>
> **Important Considerations**
>
> This score is an **estimate** based on gameplay data and should not be interpreted as a definitive measure of intelligence or cognitive ability.
>
> Factors such as familiarity with Solitaire, playing conditions, and individual strategy preferences can influence results.
>
> Your percentile ranks are based on a comparison to our research sample, which may differ from the general population.

Having a separate window that can be generated for players on demand (i.e., via player clicking of the mouse/trackpad) can help familiarize players with their scores and their meaning in a fun, engaging way and also help them better understand what goes into these calculations and why they should matter to the player.

References

Blake, C. (2020, Nov. 27). *How to measure the difficulty of a deal in Klondike-Solitaire?* StackExchange. https://boardgames.stackexchange.com/a/53516

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica*, *129*(3), 387-398. https://doi.org/10.1016/j.actpsy.2008.09.005

D'Ardenne, K., Savage, C. R., Small, D., Vainik, U., & Stoeckel, L. E. (2020). Core neuropsychological measures for obesity and diabetes trials: Initial report. *Frontiers in Psychology*, *11*, 554127. https://doi.org/10.3389/fpsyg.2020.554127

Gielis, K. (2019a). Assessment of cognitive performance in elderly life via meaningful play. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–2. https://doi.org/10.1109/ICHI.2019.8904861

Gielis, K. (2019b). Screening for mild cognitive impairment through digital biomarkers of cognitive performance in games. *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, 7–13. https://doi.org/10.1145/3341215.3356332

Gielis, K., Brito, F., Tournoy, J., & Vanden Abeele, V. (2017). Can card games be used to assess mild cognitive impairment? A study of Klondike Solitaire and cognitive functions. *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, 269–276. https://doi.org/10.1145/3130859.3131328

Gielis, K., Vanden Abeele, M.-E., Croon, R. D., Dierick, P., Ferreira-Brito, F., Van Assche, L., Verbert, K., Tournoy, J., & Vanden Abeele, V. (2021). Dissecting digital card games to yield digital biomarkers for the assessment of mild cognitive impairment: Methodological approach and exploratory study. *JMIR Serious Games*, *9*(4), e18359. https://doi.org/10.2196/18359

Gielis, K., Vanden Abeele, M.-E., Verbert, K., Tournoy, J., De Vos, M., & Vanden Abeele, V. (2021). Detecting mild cognitive impairment via digital biomarkers of cognitive performance found in klondike solitaire: A machine-learning study. *Digital Biomarkers*, *5*(1), 44–52. https://doi.org/10.1159/000514105

Gielis, K., Verbert, K., Tournoy, J., & Vanden Abeele, V. (2019). Age? It's in the game: An exploratory study on detection of cognitive aging through card games. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 651–664. https://doi.org/10.1145/3311350.3347193

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Groznik, V., & Sadikov, A. (2019). Gamification in cognitive assessment and cognitive training for mild cognitive impairment. In V. Geroimenko (Ed.), *Augmented reality games II* (pp. 179–204). Springer. https://doi.org/10.1007/978-3-030-15620-6_8

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Passell, E., Dillon, D. G., Baker, J. T., Vogel, S. C., Scheuer, L. S., Mirin, N. L., .Rutter, L. A., Pizzagalli, D. A., & Germine, L. (2019). *Digital cognitive assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) field test battery report.* https://osf.io/preprints/psyarxiv/dcszr_v1

Pedersen, M. K., Díaz, C. M. C., Wang, Q. J., Alba-Marrugo, M. A., Amidi, A., Basaiawmoit, R. V., Bergenholtz, C., Christiansen, M. H., Gajdacz, M., Hertwig, R., Ishkhanyan, B., Klyver, K., Ladegaard, N., Mathiasen, K., Parsons, C., Rafner, J., Villadsen, A. R., Wallentin, M., Zana, B., & Sherson, J. F. (2023). Measuring cognitive abilities in the wild: Validating a population-scale game-based cognitive assessment. *Cognitive Science, 47*(6), e13308. https://doi.org/10.1111/cogs.13308

Pinto, T. C. C., Machado, L., Bulgacov, T. M., Rodrigues-Júnior, A. L., Costa, M. L. G., Ximenes, R. C. C., & Sougey, E. B. (2019). Is the Montreal Cognitive Assessment (MoCA) screening superior to the Mini-Mental State Examination (MMSE) in the detection of mild cognitive impairment (MCI) and Alzheimer's Disease (AD) in the elderly? *International Psychogeriatrics*, *31*(4), 491–504. https://doi.org/10.1017/S1041610218001370

Singh, S., Strong, R. W., Jung, L., Li, F. H., Grinspoon, L., Scheuer, L. S., Passell, E. J., Martini, P., Chaytor, N., Soble, J. R., & Germine, L. (2021). The TestMyBrain digital neuropsychology toolkit: Development and psychometric characteristics. *Journal of Clinical and Experimental Neuropsychology*, *43*(8), 786–795.

Tangalos, E. G., & Petersen, R. C. (2018). Mild cognitive impairment in geriatrics. *Clinics in Geriatric Medicine*, *34*(4), 563–589. https://doi.org/10.1016/j.cger.2018.06.005

The Many Brains Project. (2024, March 10). *Featured publications*. https://www.manybrains.net/publications

Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: A comprehensive review. *Journal of the American Geriatrics Society*, *40*(9), 922–935. https://doi.org/10.1111/j.1532-5415.1992.tb01992.x

Valladares-Rodríguez, S., Pérez-Rodríguez, R., Anido-Rifón, L., & Fernández-Iglesias, M. (2016). Trends on the application of serious games to neuropsychological evaluation: A scoping review. *Journal of Biomedical Informatics*, *64*, 296–319. https://doi.org/10.1016/j.jbi.2016.10.019

Wallace, B., Goubran, R., Knoefel, F., Petriu, M., & McAvoy, A. (2014). Design of games for measurement of cognitive impairment. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 117-120). IEEE.

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*(1), 1–36. doi:10.18637/jss.v095.i01

Appendix A:
Game-Based Indicators for Healthy and MCI Samples

The information in this appendix is adopted from Gielis, Vanden Abeele, Verbert, et al. (2021).

	Healthy Group			Mild Cognitive Impairment Group				
Indicator	*M*	*SD*	*N*	*M*	*SD*	*N*	Mean Difference (Healthy - MCI)	Effect Size (Cohen's *d*)
Result-based								
Score	565.22	896.92	23	-56.30	1032.16	23	621.52	0.64
Solved (69 games total)	28			10			N/A	N/A
Game time	266107.33	100546.06	23	422283.35	243918.32	23	-156176.02	0.84
Total Moves	68.49	17.45	23	72.59	28.54	23	-4.10	0.17
Performance-Based								
Successful move percentage	95.37	4.28	23	87.45	15.86	23	7.92	0.68
Erroneous move percentage	3.65	3.62	23	6.62	6.70	23	-2.97	0.55
Rank error percentage	1.85	2.34	23	4.51	6.18	23	-2.66	0.57
Suit error percentage	2.33	2.74	23	3.59	4.83	23	-1.26	0.32
Pile move percentage	47.36	16.93	23	56.66	16.34	23	-9.30	0.56
Average cards moved	1.29	0.21	23	1.19	0.20	23	0.10	0.49
Beta error percentage	45.25	27.83	23	57.37	29.98	23	-12.12	0.42
Final beta error	0.13	0.34	23	0.33	0.47	23	-0.20	0.49
Time-Based								
Average think time	2765.71	734.83	23	4514.78	1749.75	23	-1749.07	1.30
Standard deviation think time	1999.72	812.16	23	3544.32	2181.62	23	-1544.60	0.94
Minimum think time	957.04	223.42	23	1289.55	573.65	23	-332.51	2.83

Indicator	Healthy Group			Mild Cognitive Impairment Group			Mean Difference (Healthy - MCI)	Effect Size (Cohen's *d*)
	M	*SD*	*N*	*M*	*SD*	*N*		
Average move time	722.16	169.82	23	1050.45	426.31	23	-328.29	1.01
Standard deviation move time	440.04	383.42	23	943.64	872.37	23	-503.60	0.75
Minimum move time	376.35	97.09	23	458.03	140.38	23	-81.68	0.68
Average total time	3768.38	992.82	23	5666.61	2221.33	23	-1898.23	1.10
Standard deviation total time	2560.54	1123.66	23	4191.06	2576.13	23	-1630.52	0.82
Minimum total time	741.12	234.66	23	842.41	414.25	23	-101.29	0.30
Execution-based								
Average accuracy	79.43	4.73	23	74.51	4.68	23	4.92	1.05
Standard deviation accuracy	9.74	2.67	23	10.68	2.63	23	-0.94	0.35
Minimum accuracy	51.88	18.47	23	49.06	13.72	23	2.82	0.17
Maximum accuracy	96.07	2.33	23	92.58	4.48	23	3.49	0.98
Taps	0.77	1.41	23	6.61	12.84	23	-5.84	0.64

Appendix B:
Definition of Game-Based Indicators

Table 2 is from Gielis, Vanden Abeele, Verbert, et al. (2021, p. 46).

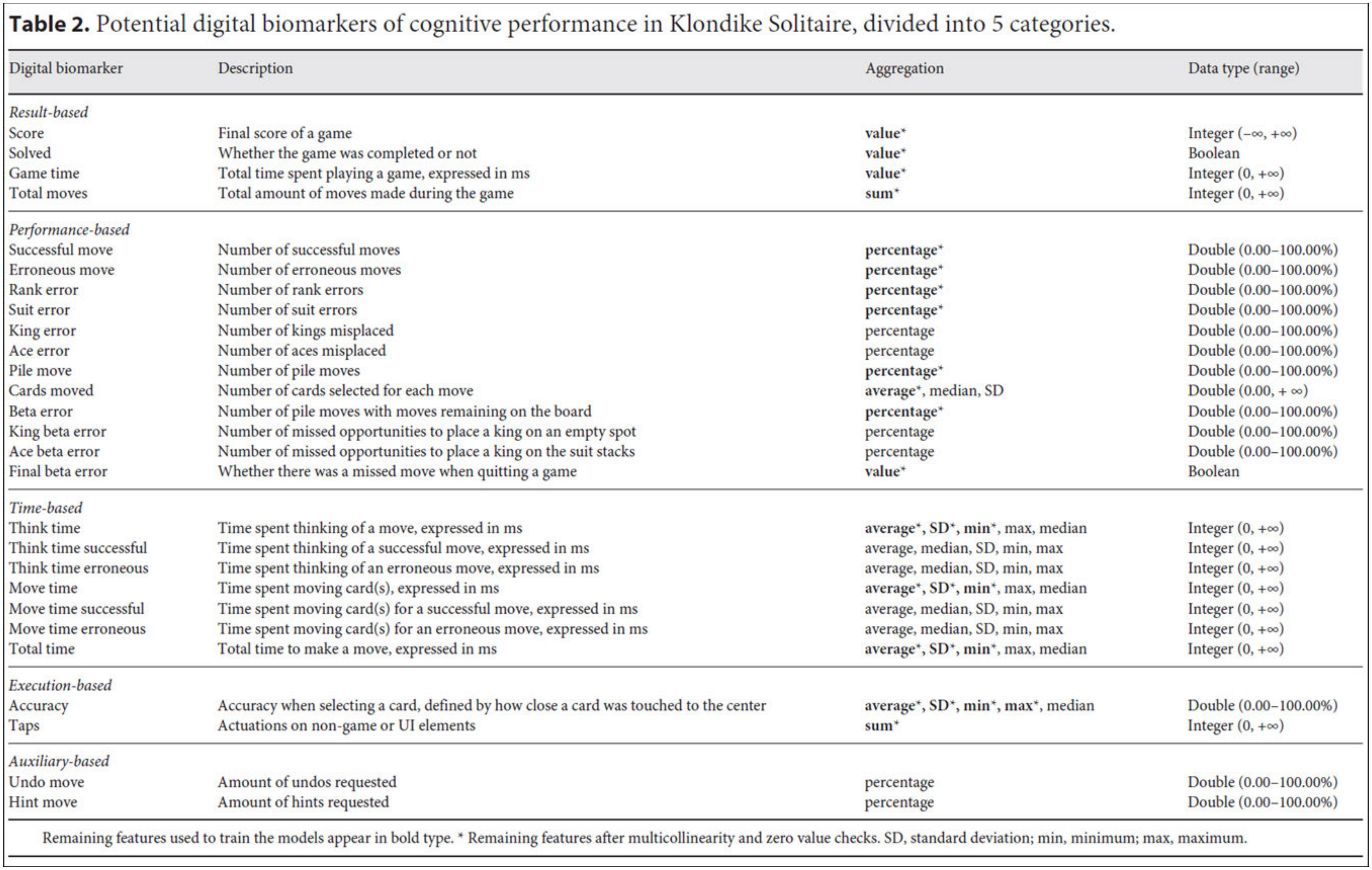


Table 2. Potential digital biomarkers of cognitive performance in Klondike Solitaire, divided into 5 categories.

Digital biomarker	Description	Aggregation	Data type (range)
Result-based			
Score	Final score of a game	**value***	Integer (−∞, +∞)
Solved	Whether the game was completed or not	**value***	Boolean
Game time	Total time spent playing a game, expressed in ms	**value***	Integer (0, +∞)
Total moves	Total amount of moves made during the game	**sum***	Integer (0, +∞)
Performance-based			
Successful move	Number of successful moves	**percentage***	Double (0.00–100.00%)
Erroneous move	Number of erroneous moves	**percentage***	Double (0.00–100.00%)
Rank error	Number of rank errors	**percentage***	Double (0.00–100.00%)
Suit error	Number of suit errors	**percentage***	Double (0.00–100.00%)
King error	Number of kings misplaced	percentage	Double (0.00–100.00%)
Ace error	Number of aces misplaced	percentage	Double (0.00–100.00%)
Pile move	Number of pile moves	**percentage***	Double (0.00–100.00%)
Cards moved	Number of cards selected for each move	**average***, median, SD	Double (0.00, + ∞)
Beta error	Number of pile moves with moves remaining on the board	**percentage***	Double (0.00–100.00%)
King beta error	Number of missed opportunities to place a king on an empty spot	percentage	Double (0.00–100.00%)
Ace beta error	Number of missed opportunities to place a king on the suit stacks	percentage	Double (0.00–100.00%)
Final beta error	Whether there was a missed move when quitting a game	**value***	Boolean
Time-based			
Think time	Time spent thinking of a move, expressed in ms	**average***, **SD***, **min***, max, median	Integer (0, +∞)
Think time successful	Time spent thinking of a successful move, expressed in ms	average, median, SD, min, max	Integer (0, +∞)
Think time erroneous	Time spent thinking of an erroneous move, expressed in ms	average, median, SD, min, max	Integer (0, +∞)
Move time	Time spent moving card(s), expressed in ms	**average***, **SD***, **min***, max, median	Integer (0, +∞)
Move time successful	Time spent moving card(s) for a successful move, expressed in ms	average, median, SD, min, max	Integer (0, +∞)
Move time erroneous	Time spent moving card(s) for an erroneous move, expressed in ms	average, median, SD, min, max	Integer (0, +∞)
Total time	Total time to make a move, expressed in ms	**average***, **SD***, **min***, max, median	Integer (0, +∞)
Execution-based			
Accuracy	Accuracy when selecting a card, defined by how close a card was touched to the center	**average***, **SD***, **min***, **max***, median	Double (0.00–100.00%)
Taps	Actuations on non-game or UI elements	**sum***	Integer (0, +∞)
Auxiliary-based			
Undo move	Amount of undos requested	percentage	Double (0.00–100.00%)
Hint move	Amount of hints requested	percentage	Double (0.00–100.00%)

Remaining features used to train the models appear in bold type. * Remaining features after multicollinearity and zero value checks. SD, standard deviation; min, minimum; max, maximum.

Appendix C:
Pilot Study Player Questionnaire

1. Age: [manual entry of number]
2. Gender
    a. Male
    b. Female
    c. Other
3. Race/ethnicity
    a. American Indian or Alaska Native
    b. Asian
    c. Black or African American
    d. Hispanic or Latino
    e. Native Hawaiian or Other Pacific Islander
    f. White
    g. Multiple races
4. Educational attainment
    a. Some high school
    b. Completed high school
    c. Some college
    d. Completed college
    e. Advanced degree
5. Do you have a cognitive impairment such as the following: mild cognitive impairment, Alzheimer's disease, traumatic brain injury, developmental disability, memory loss?
    a. Yes
    b. No
6. Are you physically active?
    a. Yes
    b. No
7. Do you have a chronic condition such as Parkinson's disease, heart disease, stroke, or diabetes?
    a. Yes
    b. No
8. Are you a smoker?
    a. Yes
    b. No
9. Do you suffer from depression?
    a. Yes
    b. No

Appendix D:
Game of the Day Win Percentage

Figure D1

Distribution of Game of the Day Win Percentages by Date

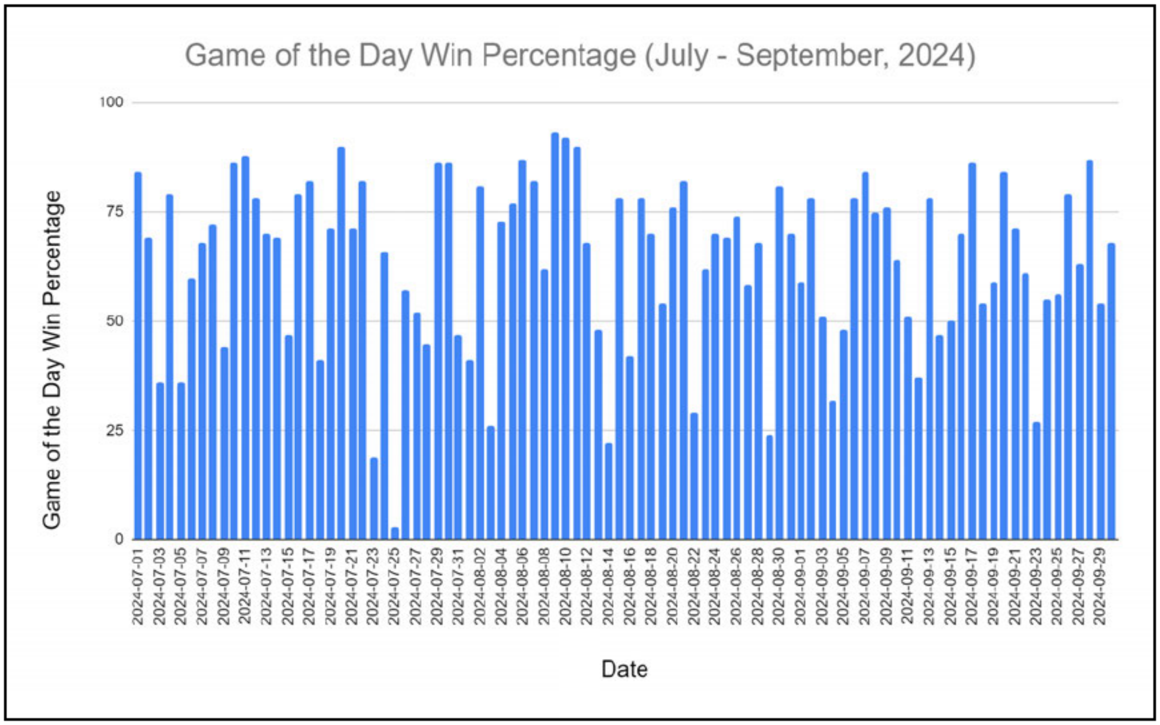


Table D1

Game of the Day Win Percentages by Date

July 2024		August 2024		September 2024	
Date	Game of the day win percentage	Date	Game of the day win percentage	Date	Game of the day win percentage
2024-07-01	84	2024-08-01	41	2024-09-01	59
2024-07-02	69	2024-08-02	81	2024-09-02	78
2024-07-03	36	2024-08-03	26	2024-09-03	51
2024-07-04	79	2024-08-04	73	2024-09-04	32
2024-07-05	36	2024-08-05	77	2024-09-05	48
2024-07-06	60	2024-08-06	87	2024-09-06	78
2024-07-07	68	2024-08-07	82	2024-09-07	84
2024-07-08	72	2024-08-08	62	2024-09-08	75
2024-07-09	44	2024-08-09	93	2024-09-09	76
2024-07-10	86	2024-08-10	92	2024-09-10	64
2024-07-11	88	2024-08-11	90	2024-09-11	51
2024-07-12	78	2024-08-12	68	2024-09-12	37
2024-07-13	70	2024-08-13	48	2024-09-13	78
2024-07-14	69	2024-08-14	22	2024-09-14	47
2024-07-15	47	2024-08-15	78	2024-09-15	50
2024-07-16	79	2024-08-16	42	2024-09-16	70
2024-07-17	82	2024-08-17	78	2024-09-17	86
2024-07-18	41	2024-08-18	70	2024-09-18	54
2024-07-19	71	2024-08-19	54	2024-09-19	59
2024-07-20	90	2024-08-20	76	2024-09-20	84
2024-07-21	71	2024-08-21	82	2024-09-21	71
2024-07-22	82	2024-08-22	29	2024-09-22	61
2024-07-23	19	2024-08-23	62	2024-09-23	27
2024-07-24	66	2024-08-24	70	2024-09-24	55
2024-07-25	3	2024-08-25	69	2024-09-25	56
2024-07-26	57	2024-08-26	74	2024-09-26	79
2024-07-27	52	2024-08-27	58	2024-09-27	63
2024-07-28	45	2024-08-28	68	2024-09-28	87
2024-07-29	86	2024-08-29	24	2024-09-29	54
2024-07-30	86	2024-08-30	81	2024-09-30	68
2024-07-31	47	2024-08-31	70		

Appendix E:
Main Study Inclusion Criteria

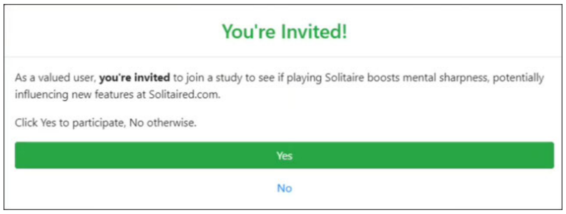
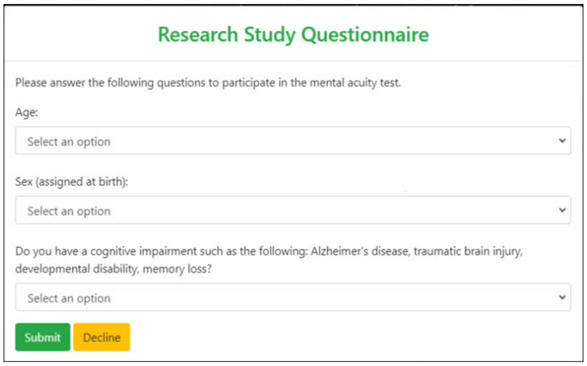
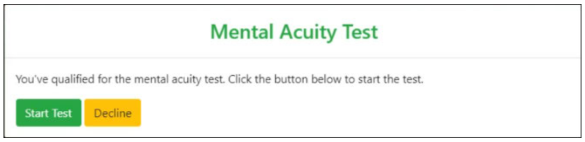
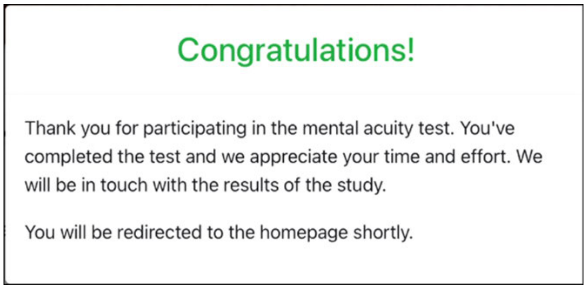
In general, the inclusion criteria represent the players and gameplay conditions that we believe will best represent the population of interest of Solitaired.com players, given practical constraints.

Criteria	Description and rationale
There are less than or equal to 100 complete players per day (i.e., wins the GoTD and completes the TMB test).	Our rationale for limiting the number of players per day is to ensure that our sample includes players who experienced different GoTDs. As shown in Appendix D, the GoTDs have a range of win percentages.
The player wins the GoTD.	Only players who win the game are included in the analyses. This strategy ensures we have complete gameplay data.
The player has not participated in the study.	We included participants in the study only once. Multiple occurrences of the same participant would violate the independence assumption (in various statistical tests) and bias the data (i.e., overrepresenting the gameplay behavior of repeating participants)
The player has not previously declined to participate in the study.	We wanted to minimize any negative perceptions of players (i.e., repeatedly prompting a player to participate when that player has declined).
The player is already registered on Solitaired.com [a]	To ensure that players were not invited to participate in the study more than once, the study is limited to registered players.

[a] Initially, only players who were existing registered players were invited to the study. This criterion was relaxed early in the data collection to increase the participation rate.

Appendix F:
Player Prompts

The study protocol contains four prompts or questions for players at several junctions. Players' responses determined how far they advanced in the protocol.

	Prompt	Screenshot of prompt
1	Invitation to participate in the study.	 You're Invited! As a valued user, **you're invited** to join a study to see if playing Solitaire boosts mental sharpness, potentially influencing new features at Solitaired.com. Click Yes to participate, No otherwise. Yes No
2	Players are asked for background information. Players can also opt out of the study.	 Research Study Questionnaire Please answer the following questions to participate in the mental acuity test. Age: Select an option Sex (assigned at birth): Select an option Do you have a cognitive impairment such as the following: Alzheimer's disease, traumatic brain injury, developmental disability, memory loss? Select an option Submit Decline
3	Players are alerted about the upcoming TMB cognitive skills test. This prompt serves as a checkpoint to prepare the player for the TMB test or allow the player to opt out of the study.	 Mental Acuity Test You've qualified for the mental acuity test. Click the button below to start the test. Start Test Decline
4	Acknowledgement after completing the TMB test.	 Congratulations! Thank you for participating in the mental acuity test. You've completed the test and we appreciate your time and effort. We will be in touch with the results of the study. You will be redirected to the homepage shortly.

Note. The prompt number corresponds to the number in the flowchart.

Appendix G:
Main Study Protocol

The study protocol is shown in Figure G1. Figure G1 is a flowchart depicting the major steps of the protocol, which involves the player and processing that occurs on Solitaired.com.

Table G1 shows the protocol from the player's point of view. Screenshots are representative of what a player would see.

The general flow begins when the player wins the game of the day (GoTD) and is invited to participate in the study. If the player accepts, then they fill out a short questionnaire and take a cognitive skills test delivered by the TMB website. After completing the cognitive skills test, the player is returned to the Soliared.com home page.

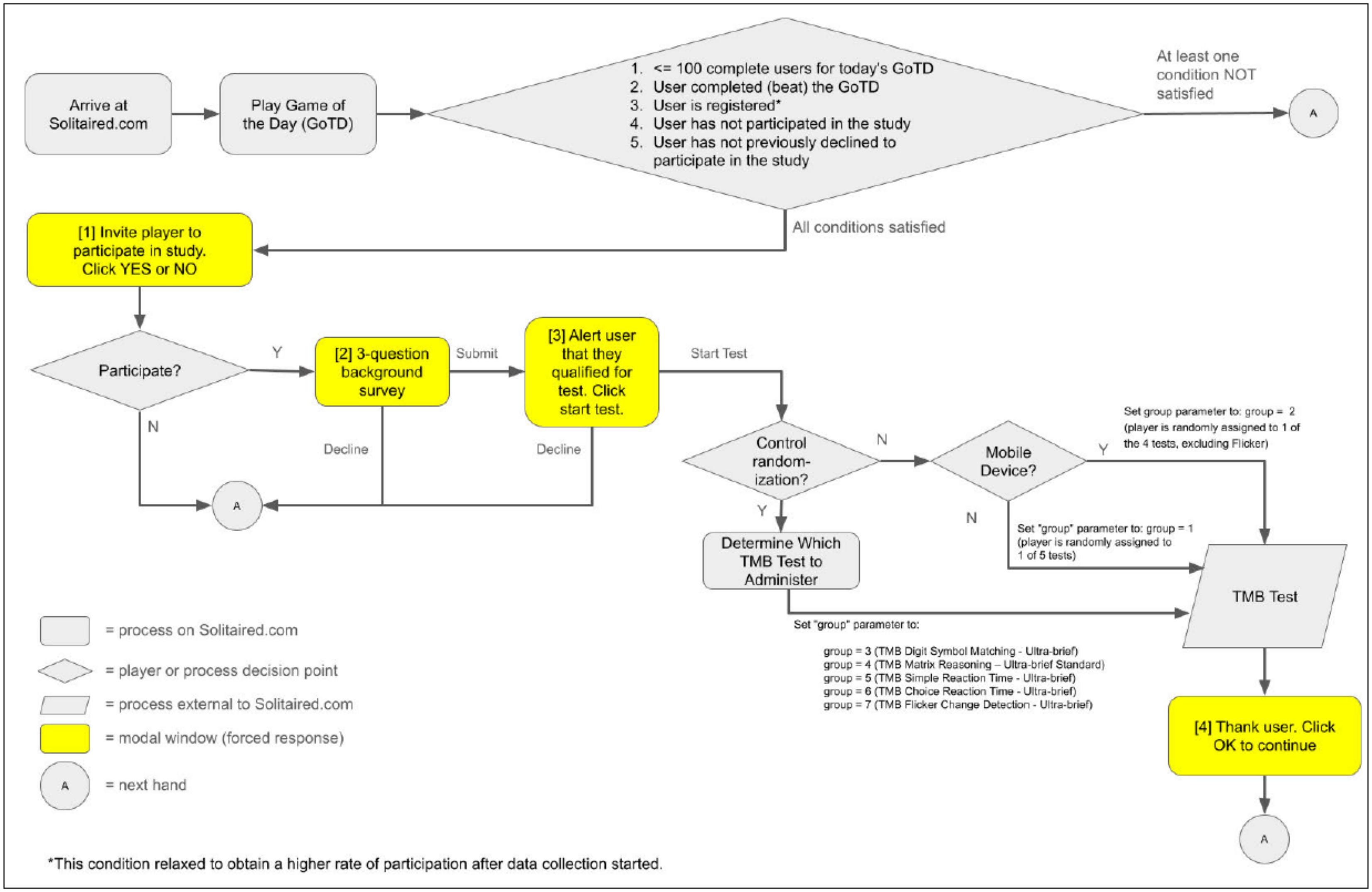
The process of answering background questions and taking the cognitive skills test should take less than 10 minutes.

The various processes and prompts shown in Figure G1 are described in greater detail in the following appendices:

- Study Inclusion Criteria: Appendix E
- Player Prompts: Appendix F
- TMB Cognitive Skills Tests: Appendix H

Figure G1

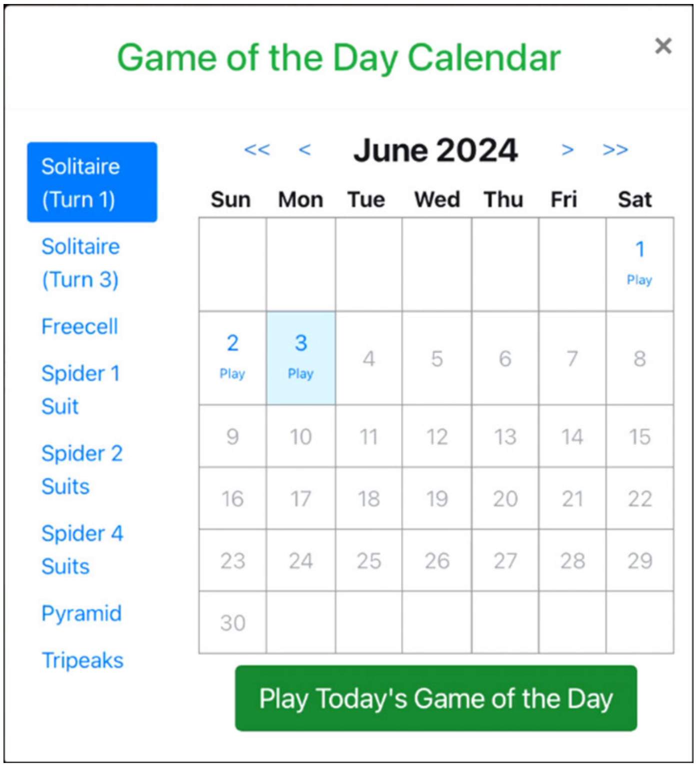
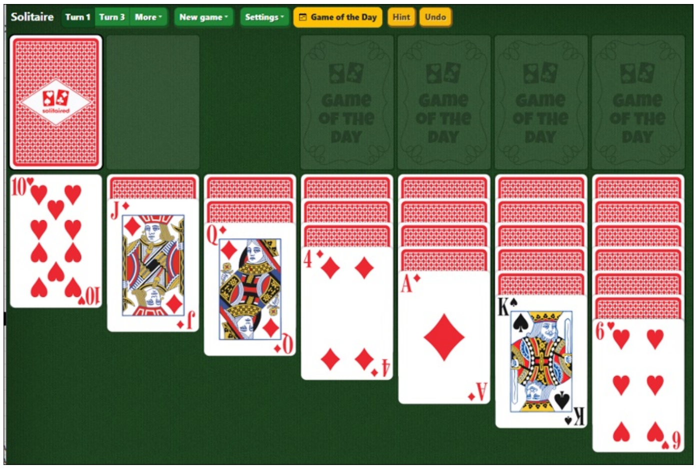
User Experience Flowchart

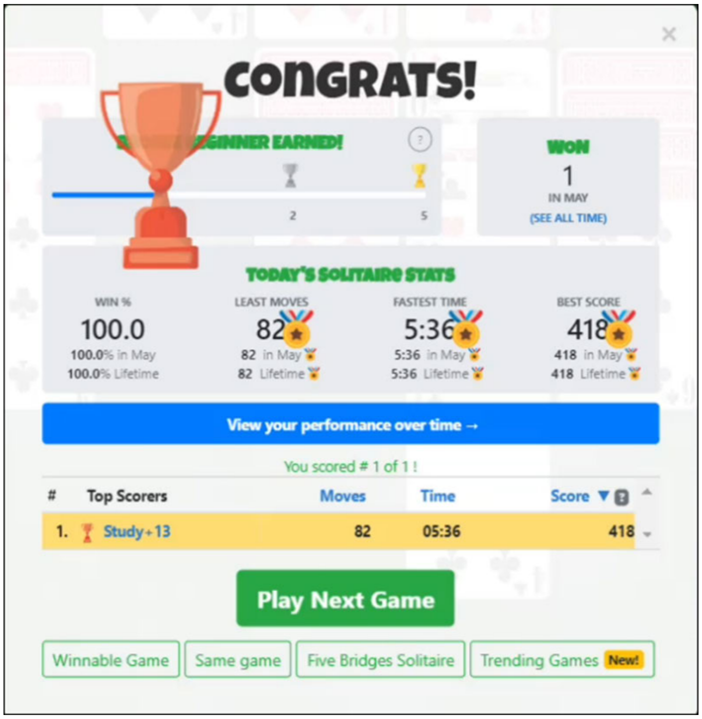
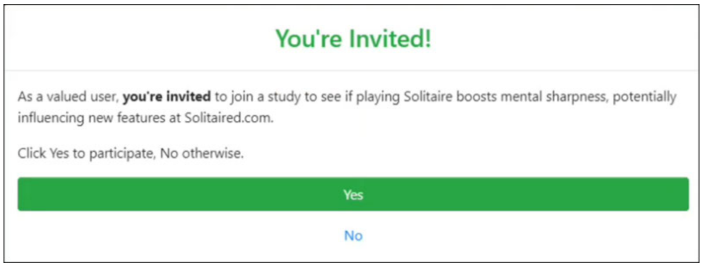
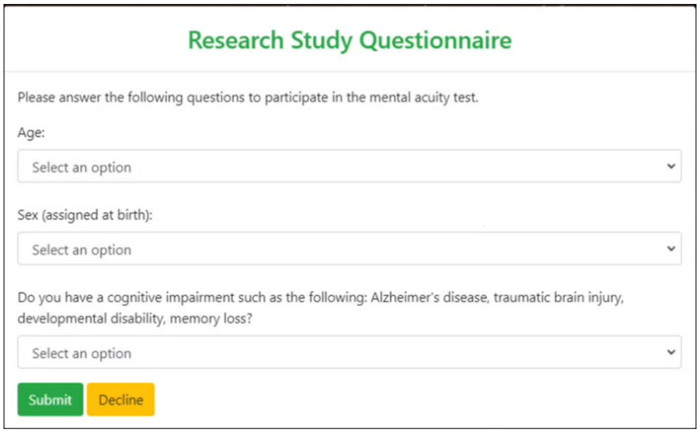
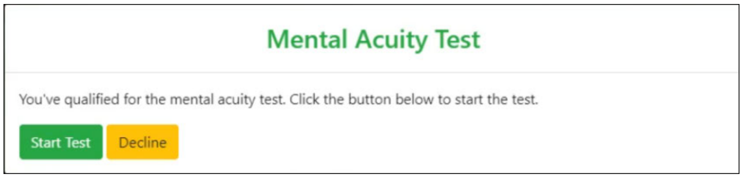

Arrive at Solitaired.com
Play Game of the Day (GoTD)
1. <= 100 complete users for today's GoTD
2. User completed (beat) the GoTD
3. User is registered*
4. User has not participated in the study
5. User has not previously declined to participate in the study
At least one condition NOT satisfied
A
All conditions satisfied
[1] Invite player to participate in study. Click YES or NO
Participate?
Y
N
[2] 3-question background survey
Submit
Decline
[3] Alert user that they qualified for test. Click start test.
Decline
A
Start Test
Control random-ization?
N
Y
Mobile Device?
Y
N
Set group parameter to: group = 2 (player is randomly assigned to 1 of the 4 tests, excluding Flicker)
Set "group" parameter to: group = 1 (player is randomly assigned to 1 of 5 tests)
Determine Which TMB Test to Administer
Set "group" parameter to:
group = 3 (TMB Digit Symbol Matching - Ultra-brief)
group = 4 (TMB Matrix Reasoning – Ultra-brief Standard)
group = 5 (TMB Simple Reaction Time - Ultra-brief)
group = 6 (TMB Choice Reaction Time - Ultra-brief)
group = 7 (TMB Flicker Change Detection - Ultra-brief)
TMB Test
[4] Thank user. Click OK to continue
A
= process on Solitaired.com
= player or process decision point
= process external to Solitaired.com
= modal window (forced response)
A
= next hand

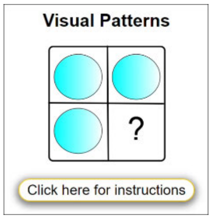
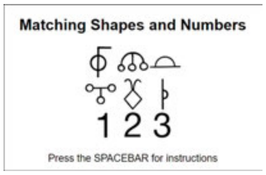
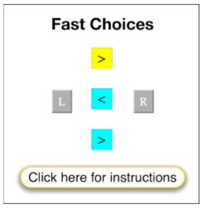
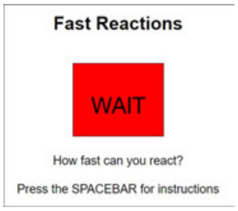
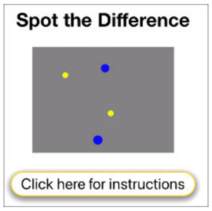
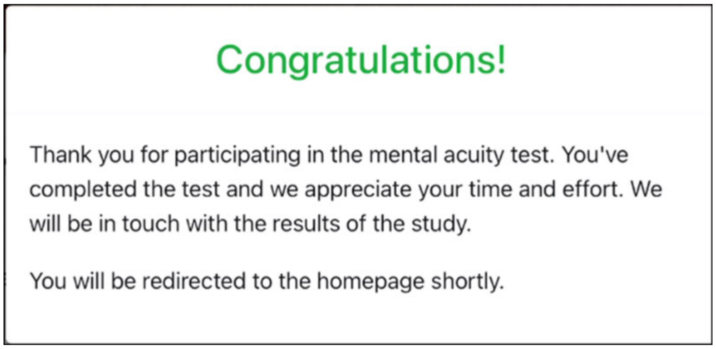
*This condition relaxed to obtain a higher rate of participation after data collection started.

Table G1

User Experience Prompts and Screens

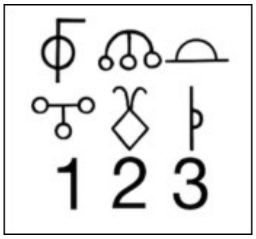
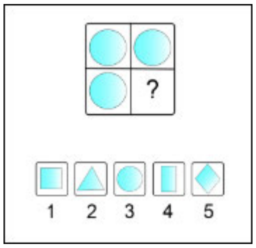
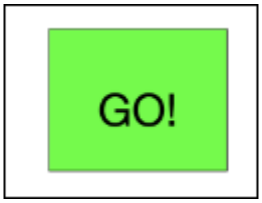
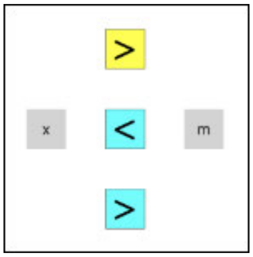
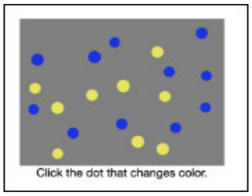
Step	User interface
In order to be included in the study, the player must play the game of the day (GoTD). Each day has a different hand of varying difficulty.	 Game of the Day Calendar × Solitaire (Turn 1) Solitaire (Turn 3) Freecell Spider 1 Suit Spider 2 Suits Spider 4 Suits Pyramid Tripeaks << < June 2024 > >> Sun Mon Tue Wed Thu Fri Sat 1 Play 2 Play, 3 Play, 4, 5, 6, 7, 8 9, 10, 11, 12, 13, 14, 15 16, 17, 18, 19, 20, 21, 22 23, 24, 25, 26, 27, 28, 29 30 Play Today's Game of the Day
This screen shows the tableau interface.	 Solitaire Turn 1 Turn 3 More New game Settings Game of the Day Hint Undo GAME OF THE DAY GAME OF THE DAY GAME OF THE DAY GAME OF THE DAY

Step	User interface
If the player wins, the *Congrats* screen is displayed. If they click on the *Play Next Game* button and satisfy the study inclusion criteria, the player will be invited to participate.	 CONGRATS! BEGINNER EARNED! 2 5 WON 1 IN MAY (SEE ALL TIME) TODAY'S SOLITAIRE STATS WIN % 100.0 100.0% in May 100.0% Lifetime LEAST MOVES 82 82 in May 82 Lifetime FASTEST TIME 5:36 5:36 in May 5:36 Lifetime BEST SCORE 418 418 in May 418 Lifetime View your performance over time → You scored # 1 of 1 ! # Top Scorers Moves Time Score ▼ 1. Study+13 82 05:36 418 Play Next Game Winnable Game Same game Five Bridges Solitaire Trending Games New!
This screen is the invitation prompt. If the player selects *No*, they will not be prompted again, and the player returns to normal gameplay. If the player selects *Yes*, they will progress to a 3-item background questionnaire.	 You're Invited! As a valued user, **you're invited** to join a study to see if playing Solitaire boosts mental sharpness, potentially influencing new features at Solitaired.com. Click Yes to participate, No otherwise. Yes No
Players are asked for their age, sex assigned at birth (male or female), and whether they have a cognitive impairment (yes or no).	 Research Study Questionnaire Please answer the following questions to participate in the mental acuity test. Age: Select an option Sex (assigned at birth): Select an option Do you have a cognitive impairment such as the following: Alzheimer's disease, traumatic brain injury, developmental disability, memory loss? Select an option Submit Decline
Players are now alerted about the upcoming TMB test.	 Mental Acuity Test You've qualified for the mental acuity test. Click the button below to start the test. Start Test Decline

Step	User interface
After hitting *Submit*, the player is randomly assigned a cognitive skills test. If the player is using a mobile device such as a smartphone, then the *Flicker Change Detection* test is excluded from the test pool, and the player is randomly assigned to one of the remaining four tests. TMB does not recommend using the *Flicker Change Detection* test on a mobile device because the screen size is too small.	Matrix Reasoning  Digit Symbol Matching  Matching Shapes and Numbers 1 2 3 Press the SPACEBAR for instructions Choice Reaction Time  Fast Choices > L < R > Click here for instructions Simple Reaction Time  Fast Reactions WAIT How fast can you react? Press the SPACEBAR for instructions Flicker Change Detection 
After completing the cognitive test, the player is thanked and returned to the Solitaired.com homepage.	 Congratulations! Thank you for participating in the mental acuity test. You've completed the test and we appreciate your time and effort. We will be in touch with the results of the study. You will be redirected to the homepage shortly.

Appendix H:
TMB Cognitive Skills Tests

The information in this appendix is adopted from TMB (2024).

TMB test	Time (min.)	Task and construct	Representative image
TMB Digit Symbol Matching - Ultra-brief	1.5	Using a symbol-number key shown on the screen, match as many symbols and numbers as possible in 90 seconds. This test measures processing speed and visual short-term memory.	
TMB Matrix Reasoning – Ultra-brief Standard	3.0	Identify the image that best completes the pattern in a series, based on a logical rule. This test has 11 items, a stopping rule, and is colorblind-friendly. This test measures fluid cognitive ability and nonverbal reasoning.	
TMB Simple Reaction Time - Ultra-brief	1.0	Press a key whenever a green square appears. This test measures basic psychomotor response speed.	
TMB Choice Reaction Time - Ultra-brief	1.0	Indicate the direction of the arrow that is a different color from the rest. This test measures processing speed, response selection/ inhibition, and attention.	
TMB Flicker Change Detection - Ultra-brief	1.0	Given a set of flashing blue and yellow dots, find the dot that is changing color from blue to yellow. This is a test of visual search, change detection, and visual working memory.	

Appendix I:
Daily Data Verification

In advance of receiving data, an RMarkdown script and an R script were developed to (a) expand the JSON fields in the Solitaired and TMB datasets; (b) check both the Solitaired and TMB data to ensure that the patterns of responses were acceptable (described below); and (c) ensure that there were no issues merging the two datasets. Initially, the data were checked every day, but starting with the fourth week, data were checked every 3 to 4 days instead of daily. The results of each data check were written to HTML documents. Variables listed below were checked.

Checking the Status of the TMB Data

Daily Tests

For each day, the total number of TMB tests, the total number at each hour, and the proportion at each hour were recorded. Ideally, tests would be distributed evenly throughout the day.

Distribution of Tests

Similarly, each day, the frequency and proportion of each of the five TMB tests were checked. Much like with the hours of the day, tests are ideally completed evenly. The mean length of time that the tests took to complete (along with a visual depiction of their distributions) were considered.

Checking the Status of the Solitaired Data

ID Uniqueness

It was important that no users were recorded in the dataset more than once. Thus, the unique user IDs in each day's data were checked against all previous days to ensure no users were recorded twice.

Daily Completes

Each day the frequency of each value in the Solitaired status column was compared. Additionally, the users with a "completed" Status in Solitaired were compared to the number of users completing the TMB test.

Registered/Unregistered Users

The proportion of each day that was made up of unregistered users was tracked/recorded.

Variability of Game Data

Each day the distributions of the Solitaired Gameplay variables were reviewed, including game completion time, hint use, undo use, and automove use. These were reviewed to identify any extreme outliers, which in turn informed the final analysis.

Variability in Survey Data

Each day the distributions of Solitaired Survey data (age, gender, self-reported cognitive impairment) were reviewed.

Appendix J:
Outlier Analysis

Removing Time Outliers

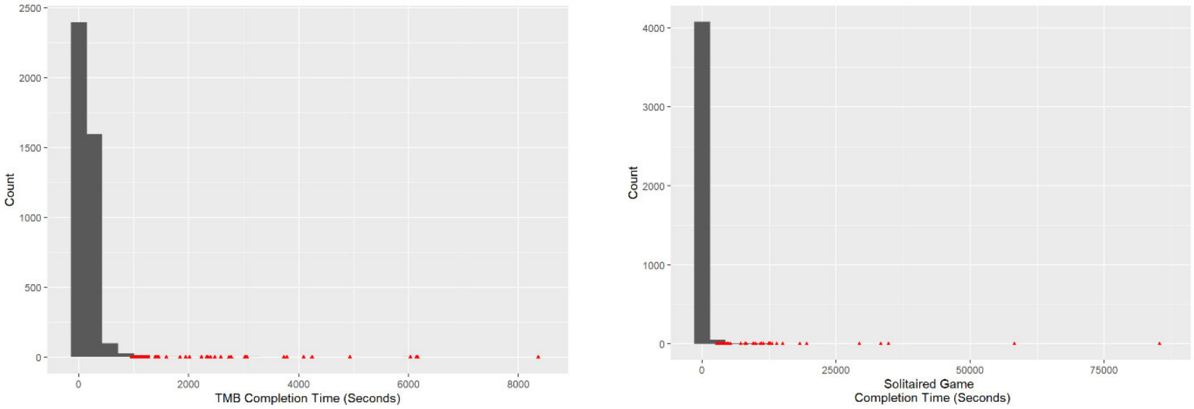
After data collection was complete, there were a total of 4,155 collected responses, ranging from 632 to 1,117 players per TMB test. However, upon initial inspection of the dataset, it was clear that two variables were particularly problematic: completion time for Solitaire games and completion time for the TMB tests (Figure J1). In both the cases, a number of players took markedly longer to complete the "Ultra Brief" tests or the Solitaire games, with completion times ranging in the hours. For instance, one Solitaired player took 24 hours to complete the game of the day, suggesting they opened the game, played it for some period of time, and then came back later to finish it. High completion times pose two major problems to our analysis:

1. They likely do not reflect the effortful play patterns we would expect from players looking to generate a mental acuity score from their Solitaired gameplay;
2. A large gap in time before finishing the Solitaire game and finishing the TMB test suggests that players may not have been in the same mental state when completing both activities, which in turn could limit the predictive power of the regression models.

Consequently, players whose times were extreme outliers were removed from the final dataset.

Figure J1

Duration (in Seconds) of TMB Tests and Solitaired Games



Removing TMB Time Outliers

Several metrics of outliers were explored, although the traditional definition is any time value that is greater than 1.5 interquartile ranges below and above the first and third quartiles, respectively (Table J1). Based on visual inspection of the above plots, we have only chosen to exclude outliers on the upper end. We can see that removing these yields a roughly 5% decrease in total sample size, which is expected and acceptable, leaving greater than 600 players for every test. Using *z*-scores yields a smaller data loss, but given the extremely skewed nature of the data it makes more sense to use the traditional definition of outlier which is robust to skewness.

Table J1

Outlier Analysis of TMB Test Completion Time for Each TMB Test

Test name	*n*	*M*	*Mdn*	*SD*	Min.	Max.	Q1	Q3	Q3 + 1.5IQR	n > Q3 + 1.5IQR	p > Q3 + 1.5IQR	+3z	n + 3z	p + 3z
Choice reaction time	635	216.05	160	414.37	82	8358	122	210	341.25	34	0.05	1403.12	6	0.01
Simple reaction time	960	82.68	67	197.62	41	6030	56	82	121	58	0.06	659.85	2	0
Digit symbol matching	811	139.53	97	328.62	44	6164	77	130	209.5	43	0.05	1082.86	6	0.01
Flicker change detection	632	196.63	172	195.61	61	3728	126	215	349.12	26	0.04	758.83	8	0.01
Matrix reasoning	1117	266.33	203	338.24	28	6144	130	301	557.5	61	0.05	1217.72	14	0.01

Removing Solitaired Time Outliers

After removing TMB outliers, we then considered Solitaired game times. Consequently, we must turn to similar outlier metrics for Solitaired games (Table J2). Using the same metric (Q3 + 1.5*IQR) we remove another 283 players, yielding a total final sample size of 3,647. The overall sample, including summaries of the other independent variables, are explored next.

Table J2

Outlier Analysis of Solitaired Game Completion Time (in Seconds)

n	*M*	*Mdn*	*SD*	Min.	Max.	Q1	Q3	Q3 + 1.5IQR	n > Q3 + 1.5IQR	p > Q3 + 1.5IQR	+3z	n + 3z	p +3 z
3930	377.16	206	1978.83	51	85289	153	296	510.5	283	0.07	6142.48	21	0.01

Appendix K:
Variable Transformation Analysis

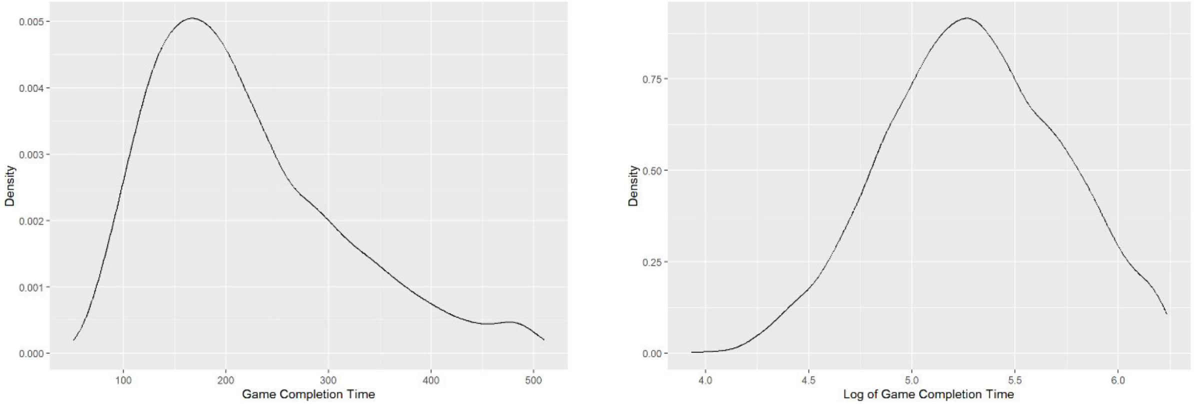
As noted in the Method section, to improve the linear regression, it was helpful to transform both the Solitaired and TMB variables to be closer to normal. Although linear regression does not have any assumptions regarding the normality of the independent or dependent variables, it does have assumptions regarding the normality of the residuals. Normalizing the independent variables is one way to potentially improve the model relative to the assumptions of linear regression. Each independent variable is explored here.

Solitaired Game Completion Time

Traditionally, right-skewed distributions can be normalized using a log transformation. Applying a log transformation to the Solitaired gameplay variable, we see the distribution becomes significantly more normal (Figure K1).

Figure K1

Distribution of Log-Transformed Solitaired Game Completion Time

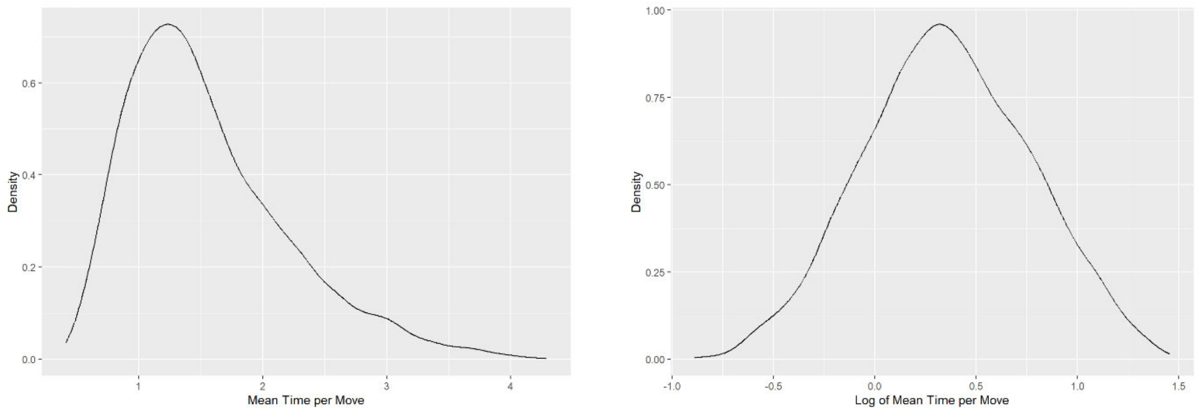


Mean Time per Move

Similar to game completion time, applying a log transformation makes the distribution more normal (Figure K2).

Figure K2

Distribution of Log-Transformed Mean Time per Move

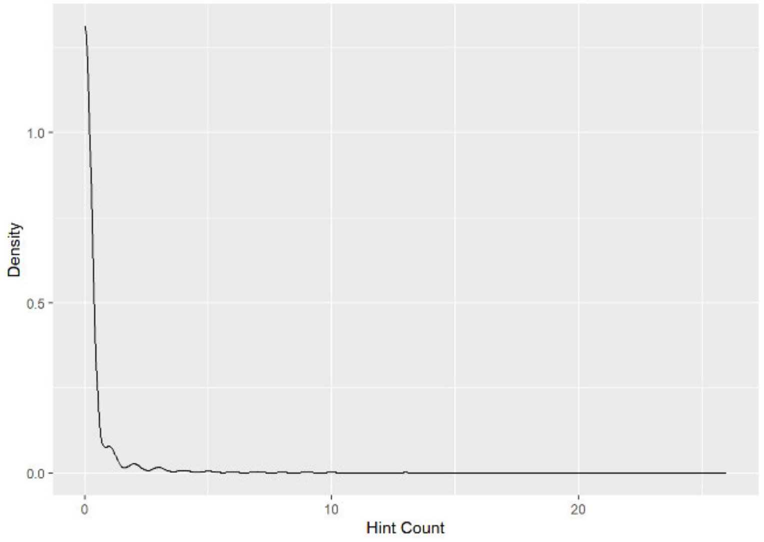


Hint Count

This distribution is incredibly right-skewed, and raises another concern altogether, which is whether this should be treated as a binary or categorical variable (Figure K3).

Figure K3

Distribution of Hint Count



As we can see from Table K1, there is strong evidence that *hint count* should be treated as binary as only a small fraction of players used hints at all. Thus, diverging from Equation 1, instead of treating this as a quantitative variable going forward it will be treated as categorical in lieu of transforming the data.

Table K1

Distribution of Hint Count

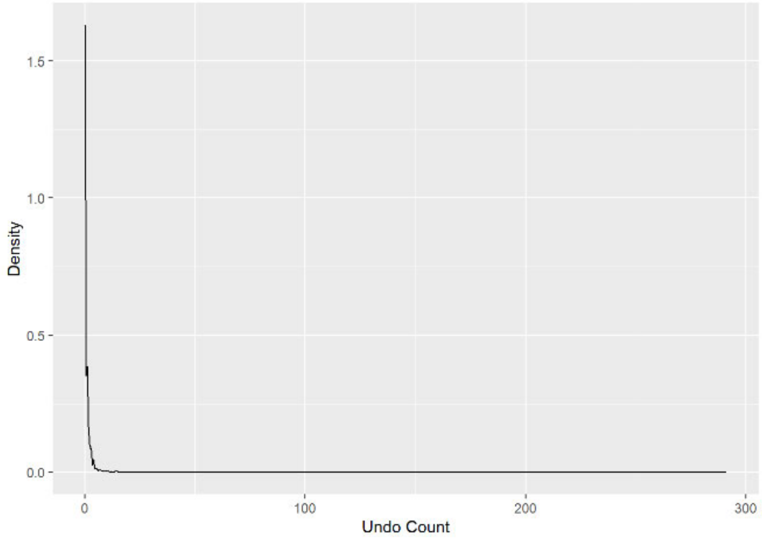
N	Min.	Max.	*Mdn*	*SD*	n > 0 hints	prop > 0 hints	n > 1 hints	prop > 1 hints
3647	0	26	0	1.54	384	.11	3647	0

Undo Count

We see an issue here similar to that for hints, once again suggesting that this variable should be treated categorically (Figure K4).

Figure K4

Distribution of Undo Count



Observing the following table (Table K2), we are once again going to treat this as categorical in the final model.

Table K2

Distribution of Undo Count

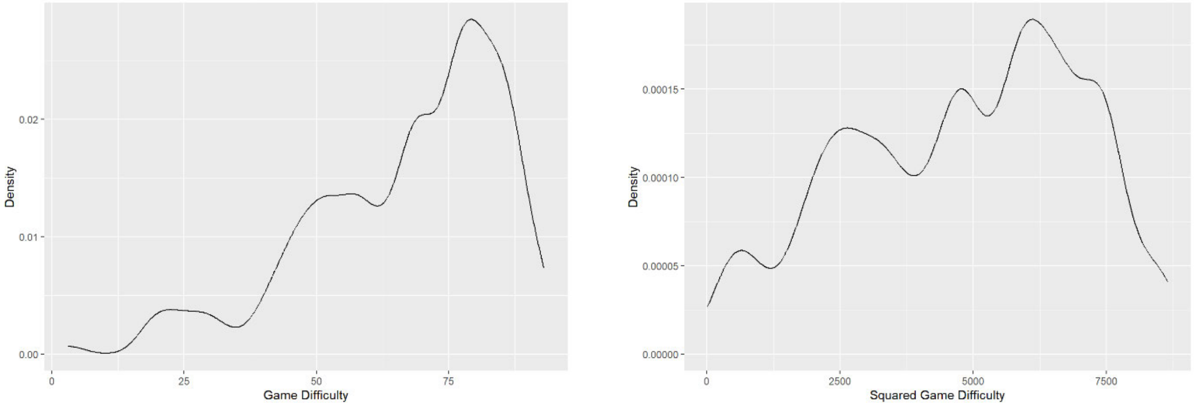
N	Min.	Max.	*Mdn*	*SD*	n > 0 undos	prop > 0 undos	n > 1 undos	prop > 1 undos
3647	0	291	0	8.57	1333	.37	602	.17

Game Difficulty

Here, we see the distribution is left-skewed (Figure K5). To account for this, we applied a squared transformation, which does appear to make the distribution slightly more normal.

Figure K5

Distribution of Game Difficulty and Squared Game Difficulty

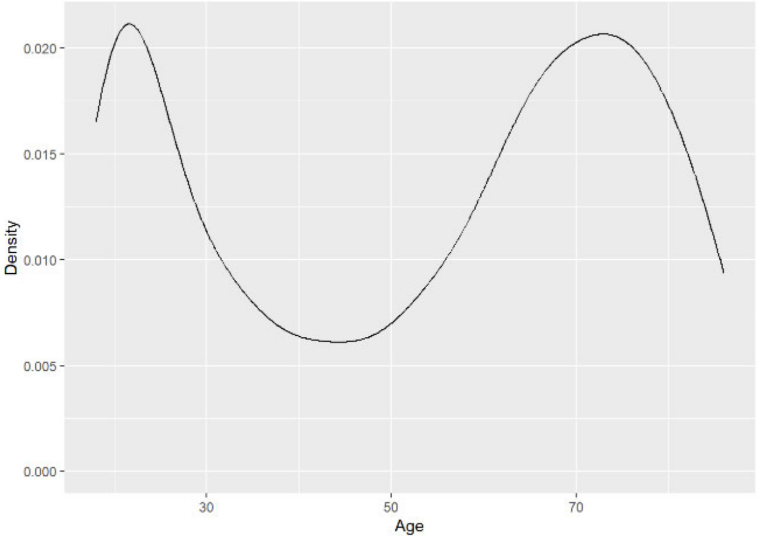


Age

This distribution is U-shaped, and there is very little that can be done to transform these data. However, as noted, there are no assumptions about the shape of the dependent variables, so this is not strictly necessary (Figure K6).

Figure K6

Distribution of Age



UCLA

CRESST

NATIONAL CENTER FOR RESEARCH ON EVALUATION,
STANDARDS, AND STUDENT TESTING

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)

School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
SE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522

(310) 206-1532
www.cresst.org

Lyons
ASSESSMENT
CONSULTING

Lyons Assessment Consulting

211 Lake Shore Drive
Wayland , MA, 01778

lyonsassessmentconsulting.com