Poster

TITLE

Validating a Machine Learning Algorithm for Measuring Cognitive Ability using Games

SHORTENED TITLE

A Gamified Measure of General Cognitive Ability

ABSTRACT

We developed six cognitive games as engaging alternatives measuring cognitive ability in the general population. A Random Forest machine learning algorithm was applied to score an array of gameplay behaviors. Gameplay score was highly correlated with general cognitive ability (r = .76, p < .01). This demonstrates games can be developed as valid measures of cognitive ability.

PRESS PARAGRAPH

Many of today's cognitive assessments, whilst digitalized, are mere adaptations of paper-and-pencil tests. We developed six short games, providing preliminary evidence that gamified assessments offer a valid alternative to conventional tests. Our research demonstrates that Artificial Intelligence techniques can be successfully applied to create valid game assessments of general cognitive ability. Our findings may encourage more companies to switch to these innovative and engaging assessments. For academics, our research details how machine learning algorithms may be used to derive scoring models for game-based ability assessments.

WORD COUNT

2954

Cognitive ability testing is widely used in applied settings, and most pervasively in recruitment and selection contexts (Bertua, Anderson, & Salgado, 2005; Schmidt & Hunter, 2004). Although most cognitive ability tests today are digitalised (Deary, 2013), the format and much of the content is simply transposed from their paper-and-pencil equivalents, with little innovation in their presentation format and the assessment experience (e.g, the Wechsler Adult Intelligence Scale, Wechsler, 1997; The Wonderlic Personnel Test, Wonderlic, 1961). More recently, gamified cognitive ability assessments have been proposed as a suitable alternative to conventional tests, as they present advantages related to both the quality and quantity of data collected, as well as the assessment experience (Lumsden, Edwards, Lawrence, Coyle & Munafo, 2016; Quiroga, Román, De La Fuente, Privado, & Colom, 2016).

First, games possess unique structural features that increase engagement and motivation in test-takers (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012). Real-time positive and negative feedback, player advancement through levels, and clear goals, are game-specific attributes that create a more motivating testing environment (Burgers, Eden, van Engelenburg, & Buningh, 2015; Wood, Griffiths, Chappell, & Davies, 2004). As such, the data collected through gamified assessments may be of better quality than that collected through more tedious pencil-and-paper formats (DeRight & Jorgensen, 2015; Miranda & Palmer, 2014). Given that motivation in high-stakes contexts is not significantly different from that of test-takers in low-stakes contexts (Sackett, Borneman, & Connelly, 2008), the positive effects of game-based assessments on motivation are likely to apply to cognitive testing in high-stakes settings such as recruiting.

Second, game-like formats show more positive reactions among test-takers. Assessments that contain gamified features tend to receive higher ratings of satisfaction and enjoyment from participants (Tso, Papagrigoriou, & Sowoidnich, 2015), but also tend to induce a heightened sense of flow (Nacke, Nacke, & Lindley, 2009; Chen, 2007; Tremblay, Bouchard, & Bouzouane, 2010).

Third, game-like tasks may elicit less anxiety for those being tested, by the reframing of the assessment as a 'challenge' (Alter, Aronson, Darley, Rodriguez, & Ruble, 2010; McPherson & Burns, 2005).

Games as Accurate Measures of Specific Cognitive Skills

Advances have been made in recent years in applying these desirable features within serious games, defined as "*games that do not have entertainment, enjoyment, or fun as their primary purpose*" (Michael & Chen, 2006, p.21). Serious games have been applied in far-reaching contexts, including education (Tobias, Fletcher & Wind, 2014), behavior change (Starks, 2014), neuropsychological assessment (Jimison, Pavel, Bissell, & McKanna, 2004), and cognitive training (O'Toole & Dennis, 2014). Serious games also possess psychometric properties that allow them to detect individual differences in cognitive ability (Quiroga et al., 2015) and to predict outcomes related to cognitive ability (Luft, Gomes, Priori & Takase, 2013).

Game-based assessments have been successfully developed for use in both assessment and educational settings, with several games validated for use in special populations, such as the elderly and children (Jimison, Pavel, Bissell, & McKanna, 2007; Jimison, Pavel, McKanna, & Pavel, 2004; Verhaegh, Fontijn, Aarts and Resing, 2013). For example, a series of three cognitive games administered via a Nintendo Wii device is moderately correlated with attention ($r = .78$, $p < .05$), memory ($r = .69$, $p < .001$), and motor control tasks ($r = .46$, $p < .001$) among an elderly population (Gamberini et al., 2010). An Artificial Neural Network successfully predicts mathematics low school achievement from a game-like cognitive task, differentiating between students with typical and low school achievement (Luft et al., 2013).

Fewer gamified assessment tools are validated for use in the general population. Gameplay across 12 separate short video games is highly correlated with cognitive ability in a sample of 188 university students ($r = .93$, $p < .01$; Quiroga et al., 2015).The Great Brain

Experiment validates a series of traditional cognitive tasks in gamified format on a smartphone (Brown et al., 2014). In a similar vein, a gamified four-dimensional spatial task predicts performance on several cognitive tasks in the general population, the highest correlations being found for a working memory span tasks (r = .62, p < .05), quantitative reasoning tasks (r = .30, p < .05), and Raven's progressive matrices (r = .37, p < .05) (Atkins et al., 2014). Gamified version of working memory and processing speed tasks also correlate with standard measures of cognitive ability, including the Wechsler Adult Intelligence scales (correlations between gameplay and cognitive ability tests between r = .54, p<.01 and r = .25, p <.05; McPherson & Burns, 2008).

Lack of Game-Based Measures of General Cognitive Ability

There is an urgent need for more extensive research that prioritises the psychometric properties of gamified assessments, and for an increased use of technology within neuropsychological and psychometric assessment (Leutner, Yearsley, Codreanu, Borenstein, & Ahmetoglu, 2017; Lumsden et al. 2016; Miller & Barr, 2017). Despite ongoing efforts to develop games with more robust psychometric properties, such efforts are constrained by their typically heterogeneous design and small sample sizes. Another key shortcoming of most game-based assessments of cognitive ability is their development for use in specific subsections of the population (e.g., children, the elderly, dementia patients). Therefore our psychometrically-focused study avoids any such preoccupation with validating cognitive ability subcomponents *only*.

In summary, our games are designed to measure general cognitive ability and so are validated against, general cognitive ability. They are developed for use with both the general UK and US populations. Furthermore, given that a wide array of behavioral and performance data is collected during gameplay in digitalized assessments, we use machine learning methods to devise a scoring model that captures the complexity of gameplay behavior.

As such, the aim of the present research is to describe the development procedure of a scoring model for predicting cognitive ability scores from gameplay behavior. The concurrent validity of the games in relation to an established pencil-and-paper test of intelligence is assessed, so that practitioners better understand how such game-based assessments compare to traditional tests. We provide a detailed description of the machine learning algorithm developed to score the game-based measure.

Method

Sample

395 participants were recruited using an online panel provided by Prolific Academic. They were rewarded for their participation. Participants downloaded the MindX mobile application to complete the six games, a traditional measure of general cognitive ability, and a demographic questionnaire. The six games were timed at 20.5 minutes, and the general cognitive ability measure at 30 minutes. 67 participants were removed from the sample due to invalid responses. They either played too few levels, or lost too many levels (bottom and top 2.5% of the sample, respectively). This resulted in valid responses from 328 participants (148 female), with an age range of 17 to 54 years (mean=32.86 years, SD=8.38).

Measures

General Cognitive Ability Measures

A total cognitive ability score was computed by combining the 16-item International Cognitive Ability Resource Sample Test (ICAR; The International Cognitive Ability Resource Team, 2014) and four items from the Cognitive Reflection Test (CRT; Frederick, 2005; Toplak, West, & Stanovich, 2014).

ICAR (The International Cognitive Ability Resource Team, 2014) is a public-domain assessment tool of cognitive ability, widely used in psychological and social scientific research

and recommend for use in high-stakes testing, including job selection and clinical purposes. ICAR has four item types: Three-Dimensional Rotation items (four items) are cube renderings to which individuals identify the response choice that is a possible rotation of the target stimulus; Letter and Number Series (four items) are short digit or letter sequences to which individuals need to identify the next position in the sequence from among six choices; Matrix Reasoning (four items) are 3x3 arrays of geometric shapes with one of the nine shapes missing, where individuals identify the geometric shape that best completes the stimuli out of six possible responses; Verbal Reasoning (four items) are logic questions.

CRT measures deliberative/rational thinking, the tendency to override a prepotent response alternative that is incorrect, and to engage in further reflection that leads to the correct response. The CRT contains scenario items such as: “In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?” [Correct answer: 47 days; Intuitive answer: 24 days]. It shows medium to high correlations with cognitive ability measures, and predicts performance on rational thinking tasks, even when cognitive ability and thinking dispositions are taken into account (Frederick, 2005; Toplak, West, & Stanovich, 2011, 2014). CRT items were thus included to provide a broader assessment of cognitive ability.

Cognitive Games

Six cognitive games, designed to be played on a mobile device, were developed to assess cognitive ability (see Table 1). The playtime of each game was fixed, and the following features were tracked for the scoring model:

- $maxCompletion = \ max\ level \div max\ level\ reachable \ \in [0,1]$
- $finalCompletion = final\ level \div max\ level\ reachable \ \in [0,1]$
- $winRatio = number\ of\ successful\ levels \div number\ of\ matches \ \in [0,1]$
- $scoreRatio = maxCompletion \times winRatio \ \in [0,1]$

$maxCompletion$ and $finalCompletion$ measure how far a player advances in relation to the maximum level reachable in the game (level 100). The $winRatio$ measures how many levels are completed correctly: If the player made no mistakes at any level, $winRatio = 1$. However, a player can have $winRatio = 1$, but have not performed the task quickly enough to play many levels in the allocated time, hence the $scoreRatio$ metric combines the $winRatio$with $maxCompletion$, yielding a normalised measure of how far and how well a player has done in the task.

Scoring Methods

Table 1 presents the machine learning models tested as scoring methods, as well as the hyperparameters used during the cross-validation procedure. 15 separate models were tested to identify the model that would best describe the relationship between gameplay and cognitive ability. Models included simple benchmarks (Mean and Median Predictors), weak-learners (Classification and Regression Trees and Linear Regression), and nonlinear models (Support Vector Regression, Multi-Layer Perceptron and Ensemble approaches) (Bishop, 2006; Duda, Hart, & Stork, 2012; and Friedman, Hastie, & Tibshirani, 2016). Each model established the following functional relationship for the subject *i*:

$$Cognitive\ Ability_i = f(Digitspan_i, Shapedance_i, Disconumbers_i, Pathfinder_i, Numerosity_i, Flashback_i) + \varepsilon_i$$

where $Cognitive\ Ability_i$is the dependent variable, $Digitspan_i, Disconumbers_i, Pathfinder_i, Numerosity_i, Flashback_i$ are the independent variables, and ε_i is the residual term. Each independent variable is decomposed in the four metrics as outlined before: $maxCompletion, finalCompletion$, $winRatio$, and $scoreRatio$.

We used a 5-fold cross-validation scheme in a nested resampling fashion. After separating the data into five training and test sets (outer level), we divided the training set into five further training and test sets (inner level). The inner level was responsible for model calibration. Using the resulting best configuration for each model, the outer level established a comparison across all models in order to identify which scoring strategy has yielded the best generalisation capacities. We repeated this process 30 times and reported the average and standard deviation of the correlation attained by each model.

Results

Table 3 shows the correlations between scores derived by the different scoring models and the cognitive ability score. Random Forest produced the highest correlations with cognitive ability, whilst the Gaussian Process Regression yielded the poorest performance. Although the Random Forest performed better than linear models traditionally used in psychometric assessment, those models still performed moderately. Some outperformed nonlinear models (Multi-Layer Perceptron and Support Vector Regressions) which tend to do well across applications in other areas.

Figure 1 shows the relative importance of each feature within the Random Forest model, in terms of variance explained. Digitspan is the most important game, with its features weighing in total around 50% of the explained variance of the model. WinRatio and *scoreRatio* are the most important features, whist the *finalCompletion* metric only seems to be relevant for Digitspan. Almost 25% of explained variance in the final prediction are accounted by the Digitspan scoreRatio. The partial dependence plot within Figure 1 shows a nonlinear relationship between the Digitspan *scoreRatio* and the game based score prediction. The graph indicates that those with *scoreRatio* below 0.47 are predicted to have below average cognitive ability scores. For those with *scoreRatio* above 0.47, an above average cognitive ability score is

predicted. The 0.47 threshold corresponds to remembering over 7 digits in the Digitspan task; the *scoreRatio* is a nonlinear predictor of cognitive ability.

Concurrent Validity

Figure 2 outlines the predicted values produced using the Random Forest model (Predicted Game Based Score) plotted against the Cognitive Ability Score. The region around 0.4 and 0.6 is where the model performs better, without underestimating (around 0.8 and 1.0) or overestimating (around 0.0 and 0.2) the measured Cognitive Ability Score of subjects. Hence, the model tended to capture the bulk of the population distribution more accurately than its tails.

Discussion

Few game-based measures of general cognitive ability exist for general populations. Where validation studies are available, they often use small sample sizes and fail to assess a broad range of cognitive abilities (Lumsden et al., 2016). We presented evidence that a series of gamified tasks are highly correlated with traditional tests of general cognitive ability. Scores on the game-based assessment correlated at $r = .76$ ($p < .01$) with general cognitive ability scores, indicating that they are a valid measure of cognitive ability.

Several gameplay behaviors predicted of cognitive ability. In addition to overall performance (*scoreRatio*), the *winRatio* was highly predictive of cognitive ability (*winRatio*). *WinRatio* may be particularly good at predicting cognitive ability because it requires maintained focussed attention. Attention forms a central part of working memory, an important component of cognitive ability (Engle & Kane, 2004). Other gameplay behaviors may be less taxing on working memory and attention because they allow the player to make mistakes.

Despite *winRatio* and *scoreRatio* being the strongest predictors of cognitive ability for most games, the strongest single predictor of cognitive ability was the final completion level for

the game Digitspan. Interestingly, the predictive power of Digitspan corresponds with research on the typical human memory capacity of seven digits (Gignac, 2015; Miller, 1956). The scoring algorithm discriminated high-performing individuals based on their ability to manipulate above seven digits. Although the relevance of working memory to cognitive ability remains subject to debate (see Ackerman, Beier, & Boyle, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005), the predictive strength of Digitspan in relation to general cognitive ability supports the theoretical and conceptual overlap between working memory and cognitive ability. This in line with research showing medium to high correlations between working memory and cognitive ability (Conway, Kane, & Engle, 2003; Kane, Hambrick, & Conway, 2005).

Nevertheless, Digitspan accounted for 50% of the shared variance between gameplay and cognitive ability scores only. This highlights the need for game based assessments to encompass several games measuring interrelated subcomponents of cognitive ability. Indeed, we found that combining gameplay features across games achieved a better prediction for general cognitive ability than individual game performance. The relatively short duration of each game makes them suitable for being used in combination with other games, thus achieving a better prediction of general cognitive ability.

Implications for Practice

Collecting psychometric data using gamified formats presents several advantages to using conventional, pencil-and-paper cognitive ability tests. Pencil-and-paper formats only allow for collection of a limited amount of data points, whereas gamified testing provides opportunities for collecting a multitude of behavioral variables (e.g., accuracy, speed, level of complexity handled by participants) in a relatively short amount of time (the games we tested lasted between 3-5 minutes each). These behavioral variables can then be used to inform the scoring of gamified assessments in order to maximize their overlap with more traditional tests of cognitive ability. Our results show that games can assess cognitive ability with similar accuracy

as traditional tests, but in a shorter timeframe. Furthermore, reduced testing times impact the ease of administering and distributing psychometric assessments on a large scale, making it suitable for collecting large quantities of high-quality data remotely.

Limitations and Future Research Directions

The finding that different gameplay behaviors were predictive of cognitive ability depending on the cognitive task, or game, performed, indicates that the separate gameplay behaviors relate differently to distinct subcomponents of cognitive ability (e.g. working memory, attention span). Future research should include conventional (i.e. pencil-and-paper) measures assessing specific subcomponents of cognitive ability in order to establish the factor structure of the gamified assessment.

Given the pervasiveness of high-stakes cognitive ability testing in applied settings, such as candidate recruitment and selection (Bertua et al., 2005; Schmidt & Hunter, 2004), the predictive validity of the gamified cognitive ability assessment needs to be established for real-life outcomes, such as job performance. This is important for validating the usefulness of gamified assessments in applied contexts, and for determining whether gamified formats provide any incremental validity in the prediction of performance, compared to more conventional psychometric formats. In a similar vein, future studies could investigate the adverse impact of gamified assessments. Population groups may adapt differently to digitalized and gamified assessment, for example, older populations may need more practice to adapt to the gamified format (Nacke et al., 2009).

Conclusion

This paper presents evidence that serious games possess psychometric properties that make them valid assessments of cognitive ability in the general population. Gameplay behavior is highly correlated with cognitive ability at $r = .76$ ($p < .001$), indicating high concurrent validity

with traditional pencil-and-paper formats of cognitive ability measurement. Therefore, such games mirror the cognitive abilities measured by conventional cognitive tests. Games may also require less time to assess cognitive ability (20.5 minutes versus 30 minutes), in part because the machine learning scoring algorithm allows for the inclusion of a wider array of behavioral and performance variables in the scoring model.

This study illustrates opportunities for innovation in the psychometric assessment of intelligence, with the potential to significantly decrease testing times and increase the quality of the data collected, thereby facilitating the measurement of individual differences in cognitive ability in the general population.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working Memory and Intelligence: The Same or Different Constructs?. *Psychological Bulletin*, *131*(1), 30-60.

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C., & Ruble, D. N. (2010). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Social Psychology*, *46*(1), 166-171.

Atkins, S. M., Sprenger, A. M., Colflesh, G. J., Briner, T. L., Buchanan, J. B., Chavis, S. E., ... & Harbison, J. I. (2014). Measuring Working Memory Is All Fun and Games: A Four-Dimensional Spatial Game Predicts Cognitive Task Performance. *Experimental Psychology*, *61*(6), 417-438.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, *78*(3), 387-409.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan, R. J. (2014). Crowdsourcing for cognitive science–the utility of smartphones. *PloS one*, *9*(7), e100662.

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*, 94-103.

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31-34.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, *59*(2), 661-686.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547-552.

Deary, I. J. (2013). Intelligence. *Current Biology, 23*(16), 673-676.

DeRight, J., & Jorgensen, R. S. (2015). I just want my research credit: frequency of suboptimal effort in a non-clinical healthy undergraduate sample. *The Clinical Neuropsychologist*, *29*(1), 101-117.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. New York, NY: John Wiley & Sons.

Engle, R. W., & Kane, M. J. (2004). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. *The Psychology of Learning and Motivation: Advances in Research and Theory*, 145–199.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25-42.

Friedman, J., Hastie, T., & Tibshirani, R. (2016). *The elements of statistical learning (2nd Ed)*. New York, NY: Springer.

Gamberini, L., Cardullo, S., Seraglia, B., & Bordin, A. (2010). Neuropsychological Testing Through a Nintendo Wii Console. *Studies in Health Technology & Informatics*, *154*, 29-33.

Gignac, G. E. (2015). The magical numbers 7 and 4 are resistant to the Flynn effect: No evidence for increases in forward or backward recall across 85 years of data. *Intelligence*, *48*, 85-95.

Jimison, H. B., Pavel, M., Bissell, P., & McKanna, J. (2007). A framework for cognitive monitoring using computer game interactions. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems* (p. 1073). IOS Press.

Jimison, H., Pavel, M., McKanna, J., & Pavel, J. (2004). Unobtrusive monitoring of computer interactions to detect cognitive status in elders. *IEEE Transactions on Information Technology in Biomedicine*, *8*(3), 248-252.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working Memory Capacity and Fluid Intelligence Are Strongly Related Constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66-71.

Kirchner, W. K. (1958). "Age differences in short-term retention of rapidly changing information.". *Journal of Experimental Psychology*. *55*(4): 352–358.

Leutner, F., Yearsley, A., Codreanu, S.-C., Borenstein, Y., & Ahmetoglu, G. (2017). From Likert scales to images: Validating a novel creativity measure with image based response scales. *Personality and Individual Differences*, *106*, 36-40.

Luft, C. D. B., Gomes, J. S., Priori, D., & Takase, E. (2013). Using online cognitive tasks to predict mathematics low school achievement. *Computers & Education*, *67*, 219-228.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. *JMIR serious games*, *4*(2), e11.

Mcpherson, J., & Burns, N. R. (2005). A speeded coding task using a computer-based mouse response. *Behavior Research Methods*, *37*(3), 538-544.

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

Michael, D. & Chen, S. (2006) *Serious games: Games that educate, train, and inform.* Boston, MA: Thomson Course Technology.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

Miller, J. B., & Barr, W. B. (2017). The Technology Crisis in Neuropsychology. *Archives of Clinical Neuropsychology*, 32, 541-554.

Miranda, A. T., & Palmer, E. M. (2014). Intrinsic motivation and attentional capture from gamelike features in a visual search task. *Behavior Research Methods*, *46*(1), 159-172.

Nacke, L. E., Nacke, A., & Lindley, C. A. (2009). Brain training for silver gamers: Effects of age and game form on effectiveness, efficiency, self-assessment, and gameplay experience. *CyberPsychology & Behavior*, *12*(5), 493-499.

Oberauer, K., Schulze, R., Wilhelm, O., & Sü, H. M. (2005). Working Memory and Intelligence—their Correlation and Their Relation. *Psychological Bulletin*, *131*(1), 61-65.

O'Toole, L. J. & Dennis, T. A. (2014). Mental health on the go: Effects of a gamified attention-bias modification mobile application in trait-anxious adults. *Clinical Psychological Science*, *2*(5), 576-590.

Quiroga, M. Á., Escorial, S., Román, F. J., Morillo, D., Jarabo, A., Privado, J., ... & Colom, R. (2015). Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can!. *Intelligence*, *53*, 1-7.

Quiroga, M. A., Román, F. J., De La Fuente, J., Privado, J., & Colom, R. (2016). The measurement of intelligence in the XXI Century using video games. *The Spanish journal of psychology*, *19*, 1-13.

Sackett, P. R., & Borneman, M. J. (2008). High-Stakes Testing in Higher Education and Employment: Appraising the Evidence for Validity and Fairness. *American Psychologist*, *63*(4), 215-227.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162-173.

Starks, K. (2014). Cognitive behavioral game design: a unified model for designing serious games. *Frontiers in Psychology*, 5(28), *1-10.*

The International Cognitive Ability Resource Team (2014). *The International Cognitive Ability Resource.* Retrieved from *https://icar-project.com*

Tobias, S., Fletcher, J. D., & Wind, A. P. Game-Based Learning. In J.M., Spector, M.D., Merrill, J. Elen, & M.J. Bishop, Handbook of Research on Educational Communications and Technology (pp.485-503). New York, NY: Springer.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20(2), 147-168.*

Tremblay, J., Bouchard, B., & Bouzouane, A. (2010, April). Adaptive Game Mechanics for Learning Purposes-Making Serious Games Playable and Fun. In *CSEDU (2)* (pp. 465-470).

Tso, L., Papagrigoriou, C., & Sowoidnich, Y. (2015). Analysis and comparison of software-tools for cognitive assessment.

Verhaegh, J., Fontijn, W. F., Aarts, E. H., & Resing, W. C. (2013). In-game assessment and training of nonverbal cognitive skills using TagTiles. Personal and Ubiquitous Computing, 17(8), 1637-1646.

Wechsler, D. (1997). Wechsler Adult Intelligence Scale-III (WAIS-III) manual. New York, NY: The Psychological Corporation.

Wonderlic, E. F. (1961). *Wonderlic personnel test manual* (No. 1). New York, NY: EF Wonderlic & Associates.

Wood, R. T., Griffiths, M. D., Chappell, D., & Davies, M. N. (2004). The structural characteristics of video games: A psycho-structural analysis. *CyberPsychology & Behavior*, *7*(1), 1-10.

Table 1

Games Used in the Study and Descriptions

Game	Duration	Description	Screenshot
Digitspan	3 min	Gamified version of the digit span sequencing task included in the widely used Wechsler Adult Intelligence Scale (Wechsler, 1997). Players memorize and reproduce (forwards and backwards) strings of letters or numbers.	
Shapedance	3.5 min	Mental rotation task that requires players to identify matching patterns of increasing complexity in rotated, and rotating, stimuli.	
Disconumbers	3 min	Players must first reproduce a sequence of numbers and then calculate the sum of the memorized numbers.	

Pathfinder	5 min	Puzzle task where players move pieces in order to create a path between two endpoints.	
Numerosity	3.5 min	Mental arithmetic task where players are presented with a set of numbers and are asked to identify a combination that when summed, subtracted, multiplied or divided, yields a target result.	
Flashback	2.5 min	Gamified version of the n-back task, measuring working memory and attention (Kirchner, 1958). Players need to decide whether the shape presented matches the one displayed 1, 2 or 3 steps back. The shapes become increasingly complex and the cognitive load of the task increases.	

Table 2.

Scoring Methods and their Hyperparameters.

Model	Abbrev.	Hyperparameters
Classic Regression	Classic Regression	None
Elastic Net Regression	Elastic Net Regression	Regularisation: {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 2.0, 3.0, 5.0, 7.0, 10.0}, and trade-off parameter: {0.1, 0.25, 0.5, 0.75, 0.9}
Gaussian Process Regression	GP Regression	Regularisation: {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 2.0, 3.0, 5.0, 7.0, 10.0}
Gradient Boosting Tree	Grad Boost Tree	Number of trees: {50, 100, 200}, and Learning rate: {0.1, 0.3, 0.5}
Lasso Regression	Lasso Regression	Regularisation: {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 2.0, 3.0, 5.0, 7.0, 10.0}
Linear Support Vector Regression	Linear SVR	Regularisation: {0.1, 0.25, 0.4, 0.75, 1.0, 2.0, 4.0, 7.0, 10.0}, and Epsilon: {0.0, 0.15, 0.3, 0.5, 1.0}
Multi-Layer Perceptron Regression	MLP Regression	Number of neurons: {5, 10, 20}, Activation function: {Rectified linear, Hyperbolic tangent}, and Regularisation: {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0}
Mean Prediction	Mean	None
Median Prediction	Median	None
Radial Basis Function Kernel Ridge Regression	RBF KRR	Regularisation: {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 2.0, 3.0, 5.0, 7.0, 10.0}, and Sigma: {0.001, 0.01, 0.1, 1.0, 10}
Radial Basis Function Support Vector Regression	RBF SVR	Regularisation: {0.1, 0.25, 0.4, 0.75, 1.0, 4.0, 10.0}, Epsilon: {0.0, 0.15, 0.3, 0.5, 1.0}, and Sigma: {0.001, 0.01, 0.1, 1.0, 10}
Random Forest	Random Forest	Number of trees: {50, 100, 200}
Regression Tree	Regression Tree	Max Depth: {1, 2, 3, 5, 7, 9}
Ridge Regression	Ridge Regression	Regularisation: {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 2.0, 3.0, 5.0, 7.0, 10.0}
k-Nearest Neighbour Regression	k-NN Regression	k: {1, 3, 5, 7, 9, 11, 13, 15, 17, 19}, and k distance metric: {Euclidean, Manhattan, Minkowski}

Table 3.

Correlations between Cognitive Ability and the Different Game-Based Assessment Scoring Models

Model	Classic Regression	Elastic Net Regression	GP Regression	Grad Boost Tree	Lasso Regression
Average	0.728	0.725	0.586	0.752	0.726
Std Deviation	0.012	0.009	0.019	0.017	0.009

Model	Linear SVR	MLP Regression	Mean	Median	RBF KRR
Average	0.723	0.729	-0.010	-0.022	0.735
Std Deviation	0.009	0.015	0.086	0.079	0.025

Model	RBF SVR	Random Forest	Regression Tree	Ridge Regression	k-NN Regression
Average	0.715	**0.757**	0.609	0.732	0.746
Std Deviation	0.018	0.011	0.038	0.009	0.017

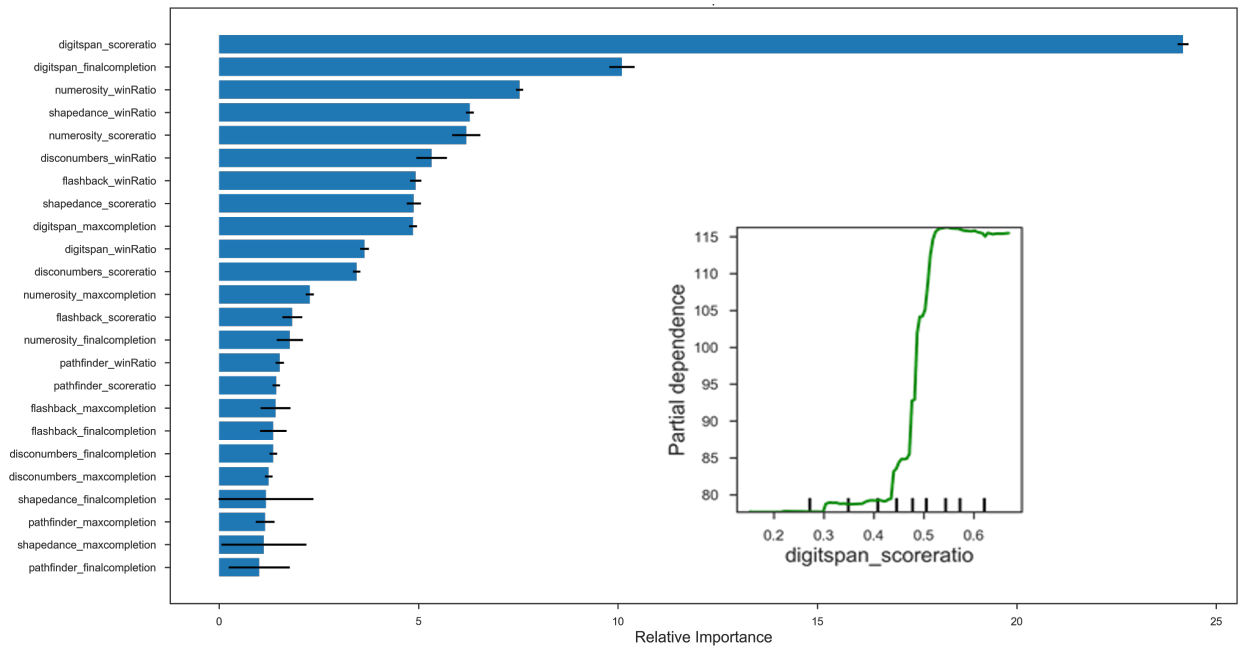

Figure 1. Feature importance (explained variance) for the Random Forest model with the partial dependence chart of Digitspan scoreRatio.

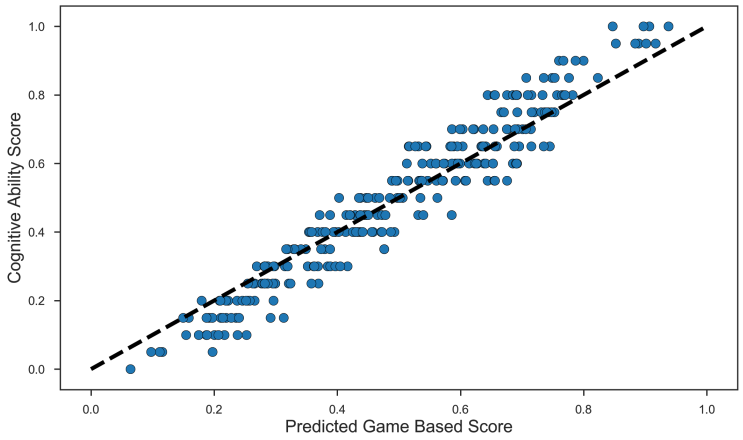

Figure 2. Measure and predicted values using the Random Forest model.