Cognitive Abilities in the Wild: Population-scale Game-Based Cognitive Assessment

Mads Kock Pedersen[1], Carlos Mauricio Castaño Díaz[1], Mario Alejandro Alba-Marrugo[2], Ali Amidi[3], Rajiv Vaid Basaiawmoit[4], Carsten Bergenholtz[1], Morten H. Christiansen[5,6,7], Miroslav Gajdacz[1], Ralph Hertwig[8], Byurakn Ishkhanyan[6,9], Kim Klyver[10,11], Nicolai Ladegaard[12], Kim Mathiasen[12], Christine Parsons[7], Janet Rafner[1], Anders Ryom Villadsen[13], Mikkel Wallentin[14], and Jacob Friis Sherson[1*]

1. Center for Hybrid Intelligence, Department of Management, Aarhus University, Aarhus, Denmark. 2. Fundación universitaria Maria Cano, Medellín, Antioquia, Colombia. 3. Department of Psychology and Behavioural Sciences, Aarhus University, Aarhus, Denmark. 4. Faculty of Natural Sciences, Aarhus University, Aarhus, Denmark. 5. Department of Psychology, Cornell University, Ithaca, New York, United States of America. 6. School of Communication and Culture, Aarhus University, Aarhus Denmark. 7. Interacting Minds Center, Aarhus University, Aarhus, Denmark. 8. Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. 9. Department of Nordic Studies and Linguistics, University of Copenhagen. 10. Department of Entrepreneurship & Relationship Management, University of Southern Denmark, Kolding, Denmark. 11. Entrepreneurship, Commercialization and Innovation Centre (ECIC), University of Adelaide, Adelaide, Australia. 12. Department of Clinical Medicine – Department of Affective Disorders, Aarhus University Hospital, Aarhus, Denmark. 13. Department of Management, Aarhus University, Aarhus, Denmark. 14. School of Communication and Culture – Cognitive Science, Aarhus University, Aarhus Denmark. *Corresponding Author: sherson@mgmt.au.dk

Psychology and the social sciences are undergoing a revolution: It has become increasingly clear that traditional lab-based experiments are challenged in capturing the full range of individual differences in cognitive abilities and behaviors across the general population. Some progress has been made toward devising measures that can be applied at scale across individuals and populations. What has been missing is a broad battery of validated tasks that can be easily deployed, used across different age ranges and social backgrounds, and in practical, clinical, and research contexts. Here, we present Skill Lab, a game-based approach affording efficient assessment of a suite of cognitive abilities. Skill Lab has been validated outside the lab in a crowdsourced broad and diverse sample, recruited in collaboration with the Danish Broadcast Company (Danmarks Radio, DR). Our game-based measures are five times faster to complete than the equivalent traditional measures and replicate previous findings on the decline of cognitive abilities with age in a large cross-sectional population sample. Finally, we provide a large open-access dataset that enables continued improvements on our work.

Keywords: cognitive abilities, gamification, crowdsourcing, individual differences

Introduction

Individual cognitive phenotyping holds the potential to revolutionize domains as wide-ranging as personalized learning, employment practices, and precision psychiatry. To get there, it will require us to rethink how we study and measure cognitive abilities. Much of what cognitive and behavioral scientists know about cognitive abilities and psychological behavior has been gleaned from studying homogeneous groups in the laboratory. Recent pushes to increase the number and diversity of participants (Bauer, 2020) are revolutionizing standards for power and generalizability across the cognitive and behavioral sciences. These advances have been enabled in part by moving from in-person testing to online equivalents, which are less costly, more convenient, and free of confounds such as experimenter expectations and testing fatigue (Birnbaum, 2004). The maturation of such tools will be critical in realizing the promise of such ambitions as individual cognitive phenotyping or precision psychiatry.

Going online with more convenient digital versions of traditional tasks makes it possible to crowdsource recruitment. Examples include projects such as LabintheWild (Reinecke & Gajos, 2015), Volunteer Science (Radford et al., 2016), and TestMyBrain (Germine et al., 2012), which offer a broad suite of digitized tasks from cognitive and behavioral science to volunteers from the general public. These scientific platforms' success in crowdsourcing data from customizable tasks has established them as a fruitful alternative to laboratory studies.

Online digital participation also opens up the possibility of developing wholly novel forms of cognitive assessment that are gamified. Gamified assessment offers the potential to engage larger and more diverse participant pools in cognitive experiments than traditional tasks and, thus, amplifies the benefits of online crowdsourcing (Baniqued et al., 2013; Lumsden, Edwards, et al., 2016). Part of the allure of adding the gamified assessment to crowdsourcing is that it motivates players by framing the activity as an entertaining and playful way to contribute to a meaningful scientific question (Jennett et al., 2014; Sagarra et al., 2016).

The gamified approach can take different directions. In one direction, the traditional task for measuring cognitive abilities is preserved as much as possible, and game-like elements, such as graphics, points, and narratives, are added to frame the task as a game. Lumsden, Skinner, et al. (2016) is an excellent example of this, where the Go/No-Go task is gamified by adding wild west illustrations and framing the task as a game, where the villains should be shoot and the innocent should be left alive. These game-like tasks have been demonstrated to be more engaging and produce similar results as their more traditional counterpart (Hawkins et al., 2013).

In another direction, new games are designed through an *evidence-centered design process* (Mislevy et al., 2003). By designing a complete game from scratch around specific cognitive abilities, researchers can obtain richer information than the traditional pen and paper version (Hagler et al., 2014). The games can be more complex and dynamic, which allows for more interesting cognitive modeling (Leduc-McNiven et al., 2018). The richer information and more complex activities could be the key to enable cognitive assessment at an individual level, rather than between groups that have been the sensitivity level of traditional tasks (Hedge et al., 2018). The cognitive assessment games often apply *stealth assessment* (Shute et al., 2016), where the cognitive ability measures are derived from the players' in-game behavior. Thus, the players

are immersed in the game experience rather than being constantly aware of being tested (Shute et al., 2016), which can help alleviate testing fatigue (Valladares-Rodríguez et al., 2016).

Prominent examples of new games built for cognitive assessment and applied at a large scale are *Sea Hero Quest* (Coughlan et al., 2019) and *The Great Brain Experiment* (H. R. Brown et al., 2014). Sea Hero Quest truly feels like a casual game experience and has reached 2.5 million participants, which yielded important insights into spatial navigation impairments in adults at risk of Alzheimer's disease (Coutrot et al., 2018). By design, Sea Hero Quest is only intended to measure spatial navigation; thus, if the goal is to measure a portfolio of distinct cognitive abilities, it would be a considerable effort to perform similar studies for each cognitive ability of interest. The Great Brain Experiment is a collection of smaller games that assess multiple cognitive abilities. Through a large-scale deployment, the games yielded new insights into age-related changes in working memory performance (McNab et al., 2015) and patterns of bias in information-seeking behavior (Hunt et al., 2016). These studies demonstrated the viability of large-scale cognitive ability testing (H. R. Brown et al., 2014) but relied on small, laboratory-based samples to validate their gamified cognitive ability measures. This raises an important question: Can we motivate large groups of players to both play the games and perform the less entertaining and more time-consuming traditional cognitive tasks in order to provide a robust within-subject validation of game-based cognitive ability measures?

Here, we present Skill Lab, an original suite of games that take advantage of online recruitment's demonstrated power to validate novel assessments of a broad portfolio of cognitive abilities. Our comprehensive mapping of multiple abilities allows us to assess their interrelations, as well as correlations with participant demographic factors, in a broad cross-section of a national population. Finally, whereas in this study, we aim to preserve a firm grounding in established theory, the benefits of the gamified approach discussed above could, in the long run – when combined with appropriate clinical tests – provide for an alternative path to *developing* a framework with a higher degree of construct and ecological validity.

Methods

In order to set up a study with the potential to contribute with new knowledge on the assessment of cognitive abilities in a crowd-focused game setup, we designated an ambitious suite of games. This process started by identifying how cognitive abilities have been operationalized and measured in laboratories. From this literature search, we selected 14 cognitive abilities (Fig 1) that we found suitable for gamification and ensures broad coverage of areas with importance for every day cognitive functioning (Lezak et al., 2012). To determine the suitability for gamfication of a cognitive ability, we had workshop sessions with game designers in which we brainstormed game-mechanics that activate the ability. The cognitive abilities we selected are related in a hierarchy where the abilities can be understood as distinct - but not necessarily orthogonal - components in different layers of cognition where more nuances are added as we move down the layers (Carroll, 1993; Deary, 2011; Jensen, 1998; Knopik et al., 2017; Mackintosh, 1998). At the top, we have a single layer representing general cognitive ability, which we subdivide into the domains of executive functioning, language, problem-solving, and visual function. The cognitive abilities that we aim to measure are sub-components of these domains.

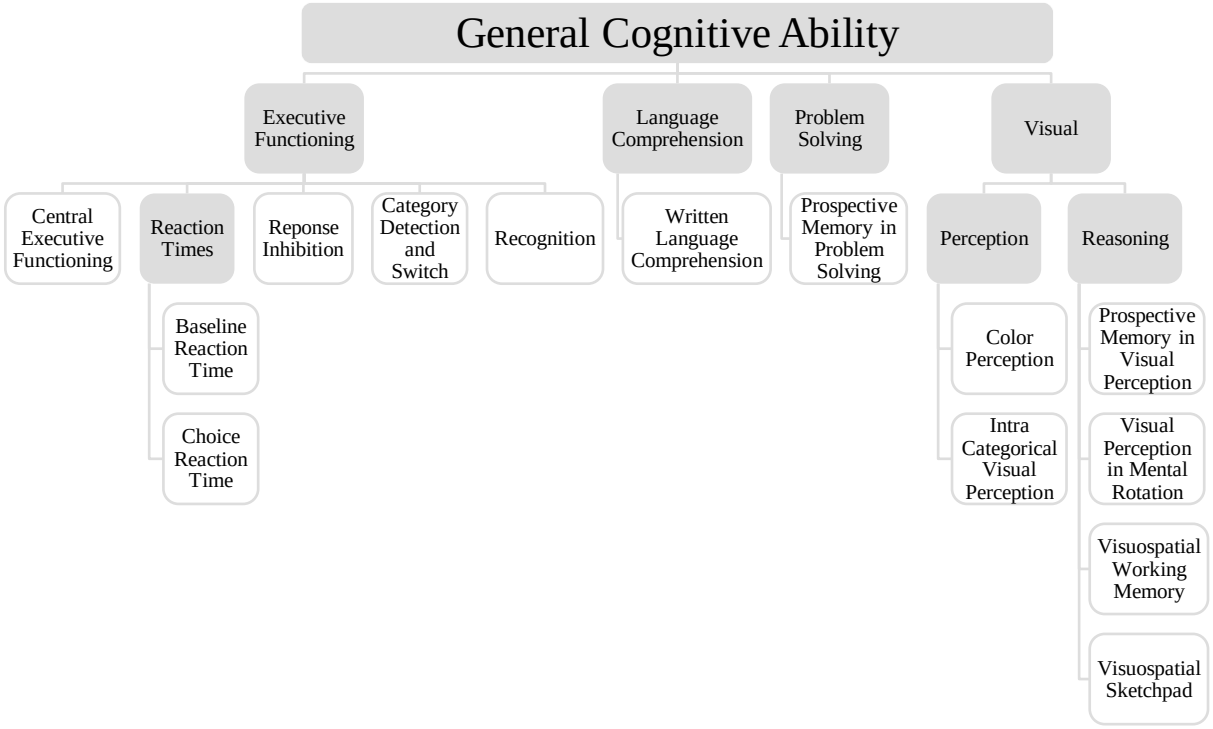

Fig 1
The 14 cognitive abilities (white boxes) that we aim to measure within the hierarcical theorethical framewordk (grey boxes). This hierarchy only expresses the relationship between the cognitive abilities that we will measure through Skill Lab, and is not a complete representation of all possible cognitive abilities, and we have not mapped all the possible relation between the components.

Game-Based Cognitive Ability Assessment

The game mechanics that were found during the brainstorming sessions were combined into the six games through the evidence-centered design process: Rat Catch, Relic Hunt, Electron Rush, Shadow Match, Robot Reboot, and Chemical Chaos (Fig 2a–f, see Supplemental Information for complete descriptions of the designs). These six games were collected into a single application called Skill Lab: Science Detective. Skill Lab contained an overarching structure and a narrative intended to motivate and guide the participant between the games. However, for this paper, we limit the scope of our analysis to the measures derived from participants' behavior within the six games.

The games were designed to measure the cognitive abilities via stealth assessment (Shute et al., 2016). We created the games with the distinctive feel of a casual game while activating the targeted cognitive abilities. A consequence of this design is that the games are not a one-to-one redesign of any particular standard cognitive task. However, there are significant shared elements allowing connections to be drawn between the cognitive abilities most likely to be activated. We could, as an example, take the relationship between the classic Go/No-Go task (Lee et al., 2009) and the Rat Catch game (Fig 2b). The Go/No-Go task, which is typically administered in test batteries, measures Response Inhibition, Baseline Reaction Time, and Choice Reaction Time (when facing distractors) by presenting a participant with a series of stimuli. If the stimulus is the correct type, the participant must react as quickly as possible;

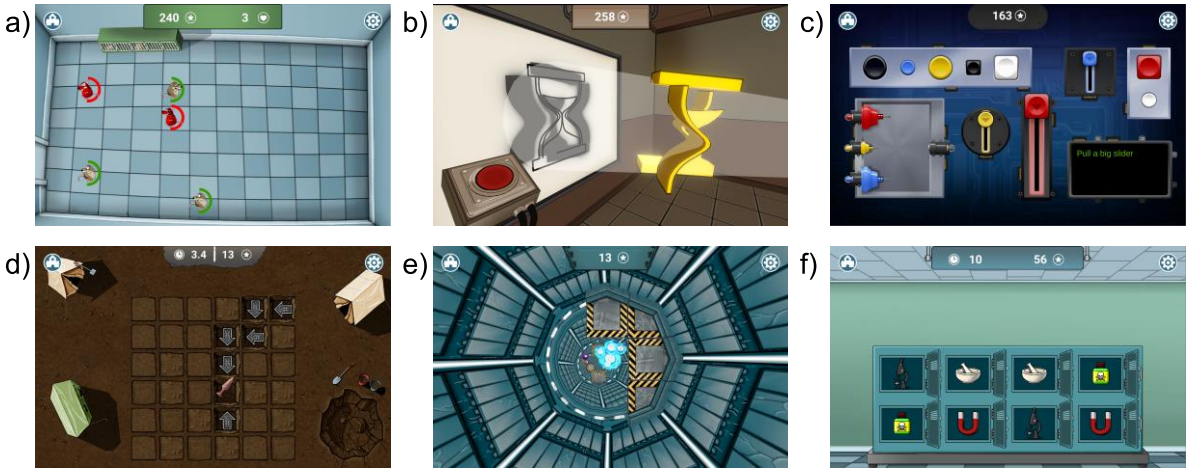

Fig 2
The six games making up Skill Lab. a) Rat Catch is designed to test Response Inhibition, Baseline Reaction Time, and Choice Reaction Time, b) Shadow Match to test visuospatial reasoning in 3D, c) Robot Reboot to test reading comprehension and instruction following, d) Relic Hunt to test visuospatial reasoning and executive functions for simple strategy making in 2D visuospatial scenarios, e) Electron Rush to test how people navigate and make decisions, and f) Chemical Chaos to measure visuospatial working memory.

otherwise, the participant should refrain from reacting. This test procedure has an analog in the first two levels of Rat Catch. In the first level, a rat appears for a limited time at a random position; the player is asked to tap the rat as quickly as possible, providing Baseline Reaction Time measures. The rats disappear faster and faster as the level progresses. Once the player misses three rats, this level of play ends.

In the second level of the game, there is a 50% chance that an "angry" red rat will appear. The player is instructed not to react to red rats but to still tap all other rats as quickly as possible. The level then follows the same progression as the first level, ending after three errors have been made (either tapping a red rat or not tapping the other rats). This taps into Choice Reaction Time and Response Inhibition. Further Rat Catch levels add variations, such as an increasing number of stimuli or moving targets that have no analog in the Go/No-Go task. These additions give indicators of visuospatial reasoning components, such as 2D spatial representation and movement perception. Through the scripted behavioral pattern assessment (Shute et al., 2016) of the game, several important game indicators and their theoretically founded relation to cognitive abilities were identified, such as average reaction time and accuracy in the different levels (see Supplemental Information).

Convergent Validity of the Game-Based Cognitive Ability Measures

Many traditional cognitive ability tasks aim to assess a single ability under strict conditions that minimize distractions and maximize experimental control (Salthouse, 2011). However, most tasks are thought to reflect multiple abilities; e.g., the Trail Making task can be used to measure Category Detection and Switch, Response Inhibition, Prospective Memory in Problem Solving, Visuospatial Sketchpad, and Central Executive Functioning (Rabin et al., 2005). In contrast, the Skill Lab games are designed to engage multiple cognitive processes, simultaneously measuring multiple abilities within a convenient, engaging, and scalable package that aims to increase ecological validity (Schmuckler, 2001) of the cognitive ability measures by creating a more realistic context and gameplay compared to traditional tasks.

To test the convergent validity of the cognitive abilities' measures from the six games, we administered 14 standard cognitive ability tasks in a separate section of Skill Lab (see Supplemental Information for full descriptions):

- Corsi Block (Kessels et al., 2000)
- Deary-Liewald (Deary et al., 2011)
- Eriksen-Flanker (Eriksen, 1995)
- Groton Maze (Papp et al., 2011)
- Mental Rotation (Ganis & Kievit, 2015)
- Go/No-Go (Lee et al., 2009)
- Stop Signal (Verbruggen & Logan, 2008)
- Stroop (Zysset et al., 2001)
- Token Test (Turkyılmaz & Belgin, 2012)
- Tower of London (Kaller et al., 2011)
- Trail Making (Fellows et al., 2017)
- Visual Pattern (L. A. Brown et al., 2006)
- Visual Search Letters (Treisman, 1977)
- Visual Search Shapes (Treisman, 1977).

To obtain quantifiable measures of the players' ability levels, we identified *indicators* of the cognitive abilities assessed (e.g., number of errors in a task) in both the games (45 indicators, see Supplemental Information) and the tasks (82 indicators, see Supplemental Information). The task indicators were found in the task literature, while the game indicators were identified through a task-analysis of the games (Newell, 1966; Newell & Simon, 1972).

As many tasks conceptually measure aspects of the same cognitive abilities, combining the observations from different tasks that have a strong theoretical overlap can give rise to composite measures of cognitive abilities that have the potential to be more robust. Measures of cognitive abilities from tasks can be defined on a spectrum of computational granularity; pure indicators (Salthouse, 2011), linear combinations of indicators (Bollen & Bauldry, 2011), all the way to methods like generative models (Guest & Martin, 2020). Here, we form linear combinations of indicators, combining indicators from multiple tasks according to a standard theoretical interpretation. We recognize that the association between any particular combination of indicators is open to debate and offer the specific aggregation of indicators here as the most straightforward theoretical proposal. (For a list of the standard tasks indicators associated with each of the 14 cognitive abilities, see Supplemental Information)

Validating Cognitive Abilities "In the Wild"

Two separate participant samples were collected for Skill Lab: i) an initial sample recruited through Amazon's Mechanical Turk (n = 444) and ii) more than 16,000 people who signed up to play the publicly available version (Fig 3a). The game was available in versions running either on mobile devices or in the browser of personal computers. Since the interactions required vary between mobile and computer versions, each would have to be separately validated (Drucker et al., 2013; Muender et al., 2019; Watson et al., 2013). Here, we focus on the mobile version since this has the broadest accessibility. Mechanical Turk's terms of service only allow data collection via the Skill Lab's browser version, so the mobile version's validation was only possible using in-the-wild participation. One of this project's contributions is a demonstration that gamified tasks' playability allows for validation with *in the wild* recruitment by motivating players to complete both games and tasks (1,351 players in Fig 3a).

Acknowledging that participant engagement typically has an exponential fall off (Lieberoth et al., 2014). A substantial player effort was needed to play both the games and complete the validation tasks; thus, broad and efficient recruitment was essential. Skill Lab was therefore launched publicly in Denmark in collaboration with the Public Danish Broadcast Company

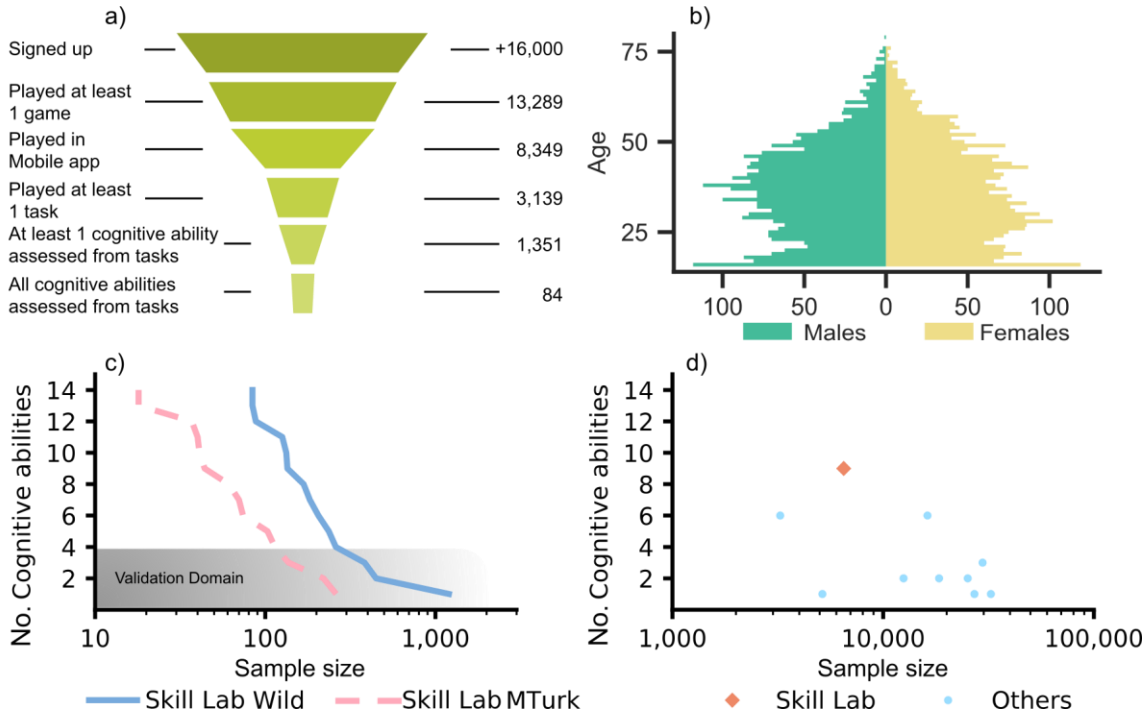

Fig 3

a) Funnel of wild player recruitment. At each layer of the funnel, fewer players had chosen to play. A small minority of players reached the bottom layer, providing enough data for us to assess all cognitive abilities from the tasks. b) Age and gender distribution for players who played at least one game in the wild. c) Simultaneous measurement of cognitive abilities from the tasks for different sample sizes from players of Skill Lab on MTurk and in the wild and the usual domain of validation (L. A. Brown et al., 2006; Deary et al., 2011; Eriksen, 1995; Fellows et al., 2017; Ganis & Kievit, 2015; Kaller et al., 2011; Kessels et al., 2000; Lee et al., 2009; Papp et al., 2011; Treisman, 1977; Turkyılmaz & Belgin, 2012; Verbruggen & Logan, 2008; Zysset et al., 2001). d) Sample size and number of cognitive abilities measured: Skill Lab games compared with other population-scale assessment studies (H. R. Brown et al., 2014; Coughlan et al., 2019; Coutrot et al., 2018; Hunt et al., 2016; McNab et al., 2015; McNab & Dolan, 2014; R. B. Rutledge et al., 2014; Robb B. Rutledge et al., 2016; Smittenaar et al., 2015; Teki et al., 2016).

(Danmarks Radio, DR), on the 4th of September 2018 on scienceathome.org, Apple Appstore, and Google Play. In Denmark, there is universal access to the internet and communication technologies (Danmarks Statistik, 2018).; To attract the broadest possible audience, we drew attention to the project through a series of DR news articles with themes varying from AI and technology to psychology and computer games (*Danmarks Nye Superhjerne - DR Retrieved: 2020-07-07 https://Www.Dr.Dk/Nyheder/Viden/Nysgerrig/Tema/Danmarks-Nye-Superhjerne*, 2020).

Results

The participants who played at least one game represent a broad cross-section of the Danish population (Danmarks Statistik, 2020) in terms of gender (5793 female, 7333 male, and 163 other; or 44%, 55%, and 1%, respectively) and age (Fig 3b), starting at age 16 years — the minimum age for granting informed consent according to the EU's General Data Protection Regulations.

Of those who played at least one game, 63% played the app version (Fig 3a). For our modeling, the players were required to have played a specific combination of tasks. We obtained a larger

sample of wild players (N=1,351) that had played a specific combination of tasks to measure a cognitive ability than from MTurk (N=444) (Fig 3c). In addition, we observed detrimental trends in parts of the MTurk data set where participants had chosen to sacrifice accuracy for speedy task completion. This trend was absent for the 'wild' players (see Supplemental Information), who chose to solve the tasks more meticulously. Given the long-term importance of the mobile platform and the success of player validation, we will present data only from the 1351 wild recruitment players using the mobile platform below, leaving the detailed comparison with MTurk and wild data for the computer platform for future studies.

Modeling Cognitive Abilities With Games

To be included in the validation process, a player had to complete at least one specific combination of tasks measuring a given cognitive ability (e.g., the three tasks Visual Pattern, Groton Maze, and Corsi Block had to be completed for us to evaluate the ability Visuospatial Working Memory).

We trained a linear model that uses game data to predict players' cognitive abilities as measured by the tasks. We started by defining cognitive ability measures by combining indicators - that measure the same construct - from different tasks. To determine which indicators to combine, we reviewed the tasks (L. A. Brown et al., 2006; Deary et al., 2011; Eriksen, 1995; Fellows et al., 2017; Ganis & Kievit, 2015; Kaller et al., 2011; Kessels et al., 2000; Lee et al., 2009; Papp et al., 2011; Treisman, 1977; Turkyılmaz & Belgin, 2012; Verbruggen & Logan, 2008; Zysset et al., 2001) and identified the indicators t_i of a cognitive ability that had a theoretical overlap (Beaujean & Benson, 2019; Mayo, 2018). For each of the 82 task indicators t_i, we assigned 14 coefficients $\alpha_{ij} \in \{-1,0,1\}$ depending on its theoretical contribution to each of the cognitive abilities C_j by assigning: 0 if there is no contribution, 1 if there is a positive correlation between the task indicator and the cognitive ability, and -1 if there is a negative correlation (see Supplemental Information for a comprehensive list of coefficients). The task indicators were standardized and combined into measures of cognitive abilities (Bollen & Bauldry, 2011) by taking weighted (α_{ij}) averages

$$C_j = \frac{\sum_{i=1}^{82} \alpha_{ij} t_i}{\sum_{i=1}^{82} |\alpha_{ij}|}.$$

For the games, we identified 45 indicators g_i from the six games that contained information pertaining to the cognitive abilities. Before any modeling was performed, all game indicators and cognitive ability measures were standardized to mean = 0 and SD = 1. Only players who had produced all the task indicators associated with the respective cognitive ability (see Supplemental Information) and at least one game indicator were included in the sample used

Table 1
Results of fitting the cognitive abilities with an elastic-net model. r_{cv} (the grey column) is the estimated out-of-sample prediction strength from the repeated cross-validation. A negative value of r_{cv} means that the model has no predictive power.

Cognitive Ability	n	r	r_{cv}	95% Confidence Interval for r_{cv}	p_{cv}
Choice Reaction Time	58	0.80	0.55	[0.34, 0.70]	0.00001
Intra Categorical Visual Perception	840	0.56	0.52	[0.47, 0.57]	<0.00001
Central Executive Functioning	185	0.63	0.52	[0.41, 0.62]	<0.00001
Baseline Reaction Time	156	0.61	0.46	[0.33, 0.58]	<0.00001
Response Inhibition	75	0.55	0.35	[0.13, 0.53]	0.00213
Visuospatial Sketchpad	131	0.52	0.32	[0.16, 0.47]	0.00015
Category Detection and Switch	88	0.6	0.28	[0.07, 0.46]	0.00771
Visuospatial Working Memory	197	0.44	0.27	[0.13, 0.39]	0.00016
Visual Perception in Mental Rotation	314	0.44	0.26	[0.15, 0.36]	<0.00001
Prospective Memory in Problem Solving	117	0.51	0.24	[0.06, 0.40]	0.01011
Prospective Memory in Mental Rotation	308	0.36	0.15	[0.03, 0.25]	0.00836
Color Perception	289	0.30	0.14	[0.02, 0.25]	0.01644
Recognition	160	0.35	0.04	[-0.11, 0.19]	0.62412
Written Language Comprehension	193	0.28	-0.02	[-0.16, 0.12]	0.74664

to fit the linear regression models predicting the cognitive abilities measured from the tasks with game indicators (for sample sizes see Table 1). Any missing game indicators were imputed using multivariate imputation with chained equations (Buuren & Groothuis-Oudshoorn, 2011), which generated one common imputation model for the entire data set. The imputation model was generated from game indicators only and contained no information about task indicators or demographic information. To prevent overfitting, an elastic-net model

$$C_j = \sum_i \beta_{ij} g_i + k_j$$

was fitted using 100 times repeated 5-fold cross-validation (Burman, 1989). If a single game indicator or the cognitive ability measured by tasks was more than 3 SD's from the mean, the player was excluded from the fitting of that specific cognitive ability's prediction model, as the fitting would be sensitive to such outliers. The elastic-net model avoids overfitting by performing variable selection and shrinking and mitigates multicollinearity issues (Zou & Hastie, 2005). The trained models ($\{\beta_{1j},...,\beta_{45j}\}$, k_j) (see Supplemental Information) are the result of averaging all the 500 individually trained models per cognitive ability. To remove bias due to overfitting to the data from the full models' correlation with the tasks (r, Table 1) we estimated an *out-of-sample prediction strength* ($r_{cv,}$ Table 1), i.e., what the correlation between the model predicted and the task measured cognitive abilities would be in an entirely new

dataset. The estimate is the average correlation between the model predictions and the task-measured cognitive abilities on the test samples for each of the repeated cross-validation test sets. The fitting and cross-validation process resulted in 10 accepted (r_{cv} >0.2) prediction models with medium to strong effect sizes and four rejected models (Table 1).

Predictive Power of Main Factor

Since we take the abilities to be hierarchically related (Fig 1), it is essential to distinguish between shared variation at different levels of the hierarchy, contributing to the observed predictive power. Therefore, we performed an exploratory factor analysis with principal factor extraction and no rotation on the cognitive abilities from both the tasks (N = 80) and the games (N = 6,369) to identify the main factor in both sets, interpretable as a generalized cognitive ability (Knopik et al., 2017). The factor analysis's exclusion criterion was whether the cognitive ability measure was more than 3SD's from the population mean. This criterion was different from the one applied during the fitting procedure, as a single outlier among the game indicators could potentially be compensated for in the predictive model, either by all the other non-outliers or that a particular game indicator is irrelevant for that particular model. Thus, we decided to exclude based on the predicted value rather than at the game indicator level. The same criterion is used for all the following analyses in this paper. For the cognitive abilities measured by the games, that means 976 (1%) cognitive abilities were excluded from 388 (5%) players. For the factor analysis, only the 6,369 players with no excluded cognitive abilities from the games were included in the analysis, and 80 out of the 85 players with all cognitive abilities measured by the tasks were included. The relatively low task participant number reflects that ten cognitive ability tasks were required to be included in the analysis.

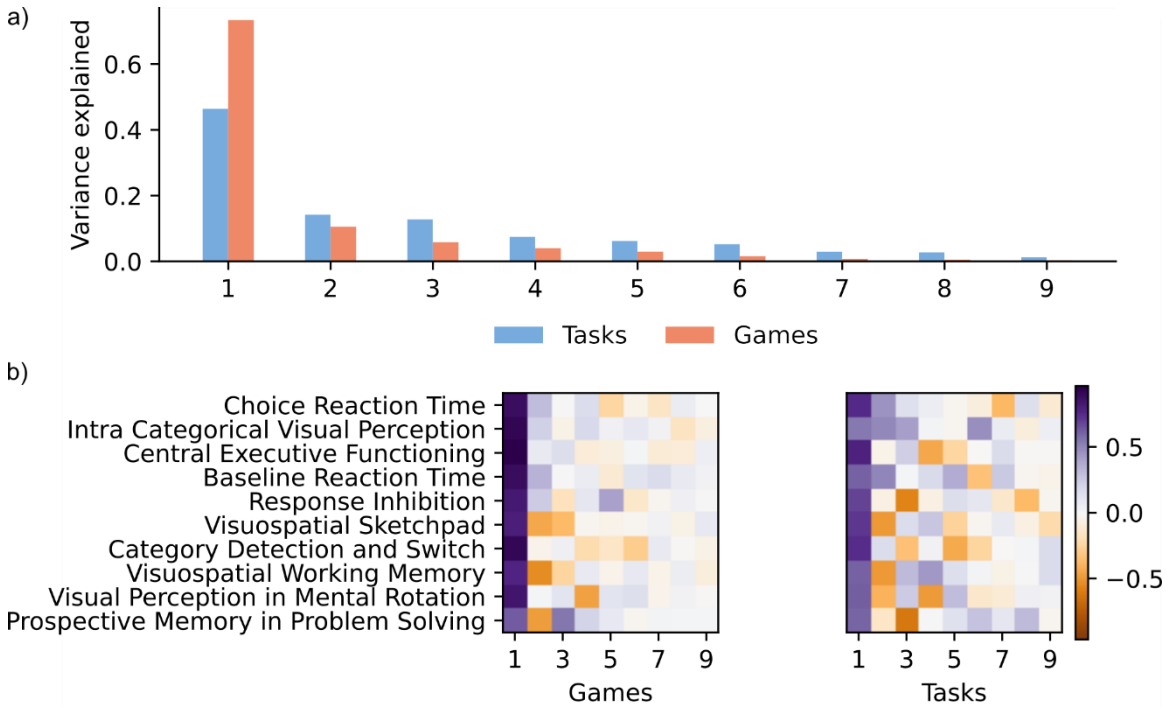

Fig 4
a) Proportion of variance covered by each factor. b) Loadings of each cognitive ability on the factors.

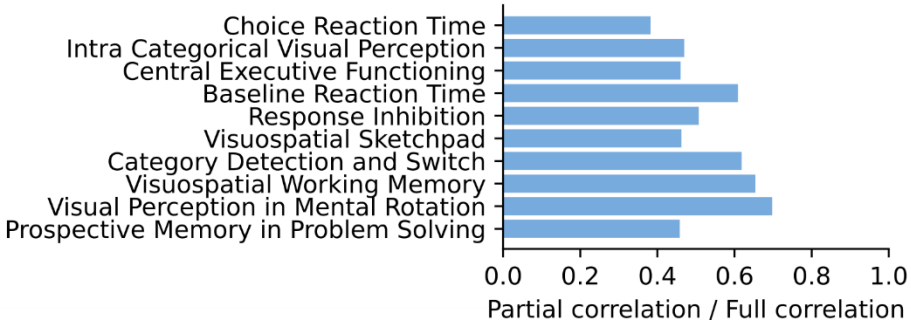

Fig 5
The proportion of the models' predictive strength not explained by the main factor. The full corrlations are similar to, but not exactly equal to, the r values found in the Table 1. A table with values of the full and partial correlation can be found in the Suplementary materials.

The components in the hierarchical framework are not orthogonal, and unsurprisingly they all tap into the general cognitive ability. However, the subsequent factors yielded by the factor analysis (principal factor extraction with no rotation) have to be orthogonal to the main factor and would not uncover the hierarchy of our theoretical framework. We are only interested in the main factor that explains the most variance when evaluating the general cognitive ability level's role.

The variance explained by the main factor increased from 46% to 72% (Fig 4a) from the tasks to the games. This was expected since the number of indicators used to evaluate the cognitive abilities had decreased from 82 task indicators to 45 game indicators. Since all the indicators are standardized to have zero mean and variance equal to 1 there is less overall variance to be explained, yielding a higher proportion of the variance explained by the main factor.

The main factor loadings are very similar for the tasks and the games and correspond approximately to the mean of the cognitive abilities (Fig 4b). In order to tell whether the main factor is the driver of the models' predictive power, we computed partial correlations between the tasks and the games while controlling for the games' main factor. These partial correlations reveal to what extent the models predict the nuances contained within each cognitive ability that goes beyond the individuals' overall ability level. We are thus interested in what fraction of the correlation between the task and the game-based measures that is not explained by the main factor. Thus, we divide the partial correlations with the full correlations (Fig 5), and we find that at least 38% of the correlation is not due to the main factor for all the models.

Temporal Stability of Models

One of Skill Lab's potential use cases is as a low cost test battery that could be used to track cognitive impairments. We are therefore interested in the stability of the cognitive ability measures, in the sense that they are reproduceable within a manageable time frame. We assessed the sensitivity (correlation between game-predicted and task-measured) and stability (correlation between playthroughs) of cognitive ability measures over repeat playthroughs (Fig 6). Only participants with at least four playthroughs were included to avoid confounding cohort effects due to sample biases. Including all participants produce similar results, indicating little to no sample bias in the temporal drift.

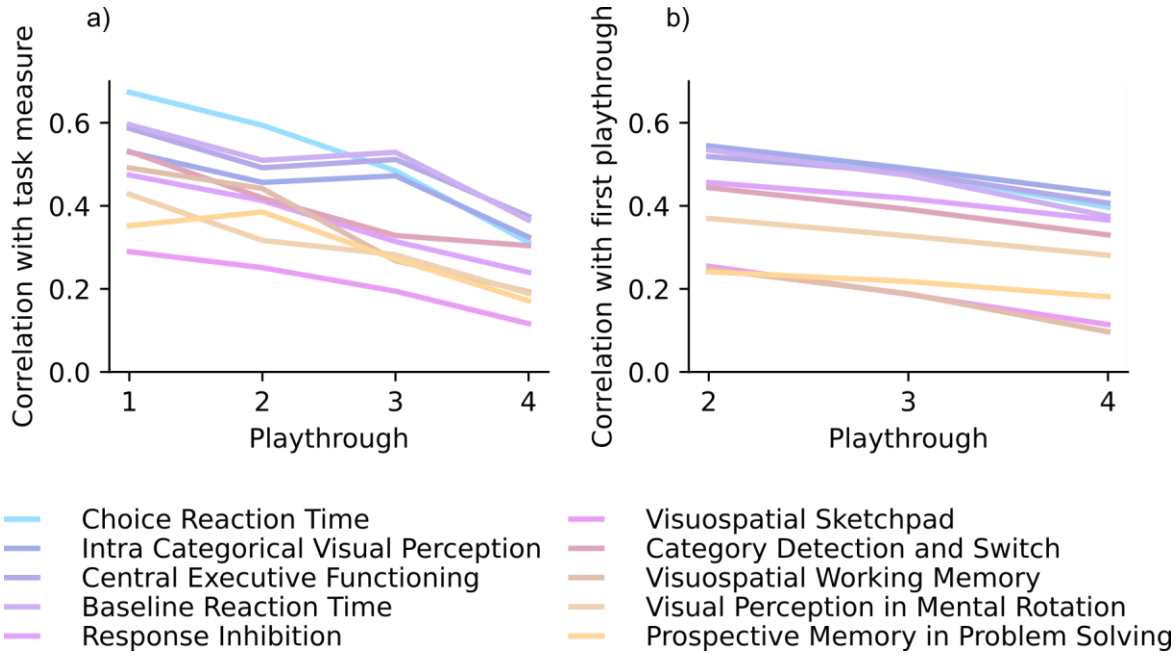

Fig 6
a) Temporal sensitivity. The Pearson correlation between the game predicted and the task measured cognitive abilities. b) Temporal stability. The Pearson correlation between the game predicted cognitive abilities between the later playthroughs and the first playthrough.

The sensitivity (Fig 6a) declines with playthroughs as expected. It should be noted that this correlation is the full correlation (not the out-of-sample prediction strength r_{cv}), as it does not account for overfitting; thus, it will be systematically higher and should primarily be used to compare across playthroughs. The second playthrough cognitive ability estimates correlate reasonably well with those from the first (Fig 6b). Throughout four playthroughs, the correlations with the first playthrough decline. As the data for this repeated play analysis comes from a convenience sample, we could not control how much time there was between playthroughs, which could also vary for the individual games in a playthrough. We can estimate the time between playthroughs by taking the first game's start as the beginning of a playthrough. We note that most repeat plays were on the same day, with three-quarters of all return plays occurring within four hours. Thus, we acknowledge that this playthrough comparison does not have the same internal validity as a traditional lab-based test-retests study.

Time Requirement for Participation

If Skill Lab is to be used as a low cost self-administered alternative to current cognitive batteries, one of the relevant parameters to look at is the time it takes to obtain estimates of the cognitive abilities. In our case we can compare the time it takes to complete all the games combined (Mean:14 min, SD: 5 min) and all the cognitive ability tasks combined (Mean:72 min, SD:7 min).

Differences in Cognitive Abilities Across Age

We use the trained models to illustrate the cross-sectional cohort distributions of cognitive abilities by age for the Danish population (Fig 7 and Fig SI. 48-57). The age bins were generated by requiring a minimum of 30 people in each bin — large enough to show differences between each bin, but small enough for at least two bins to be generated for the curves extracted

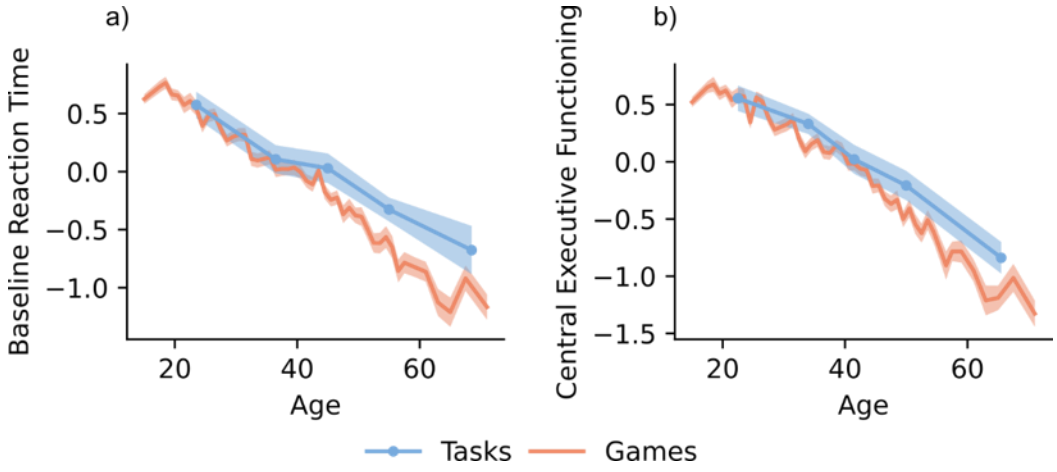

Fig 7
Cognitive abilities across age groups. 6369 wild players played the games; fewer played the combination of tasks that allowed for assessing a specific ability. The shaded areas around the curves are the standard error of the mean. Each age point in the graph includes at least 30 players (the curves for the remaining cognitive abilities can be found in the Supplemental Information). The grey lines indicate where the population crosses zero. b) Central Executive Functioning (n_{task} = 257), a) Baseline Reaction Time (n_{task} = 230).

from the task-measured cognitive abilities. The points were generated by starting at age 16 and checking whether 30 players of that age whose data provided a cognitive ability measure. If there were enough players, the next point was generated starting with those 1 year older; if not, the following ages were added 1 year at a time until a sample size of 30 was reached. Examining the distributions obtained from the games across ages, we observed the expected increase in all cognitive abilities from age 16 to 20 years, followed by a gradual decline from age 20 years.

Discussion

The combination of sample size and breadth of cognitive abilities measured in Skill Lab is exemplary (see the orange diamond in Fig. 2d) relative to other game-based population-scale assessment studies such as SeaHero Quest and The Great Brain Experiment (H. R. Brown et al., 2014; Coughlan et al., 2019; Coutrot et al., 2018; Hunt et al., 2016; McNab et al., 2015; McNab & Dolan, 2014; R. B. Rutledge et al., 2014; Robb B. Rutledge et al., 2016; Smittenaar et al., 2015; Teki et al., 2016); 6,369 players generated sufficent data for the ten trained models to be applied. In this paper, we have worked to establish the construct validity of our measures. Fig 1Ten of the models predicting the cognitive abilities from game indicators correlate well with the task-based measures demonstrating good convergent validity in line with the goals of our design process. The factor analysis revealed a main factor for cognitive abilities that could be interpreted as a general cognitive ability for both games and tasks (Fig 4) in line with a priori expectations (Fig 1). Via partial correlations (Fig 5) we demonstrated that the shared information from the main factor is insufficient to explain a substantial proportion of each cognitive ability's observed agreement between task and game estimates. Each of our measures, therefore, captures some of the nuances of the cognitive abilities beyond the dominant factor. We defer to future work a closer examination of the relationship between these quantities and application-specific standard measures. Although it is possible that more advanced modeling of the existing data set can improve these results, the ten accepted models already represent a broad, strong, and rapid testing battery.

The cognitive ability estimates are reasonably stable over repeat playgiven the short interaction time.This stabilityis encouraging for one-shot assessment applicationssuch as screening for cognitive impairments. We take the close relationship between estimates derived from first and second playthroughs as a test of how effectively the elastic-net could avoid overfitting and give well-calibrated estimates of out-of-sample prediction strengths. For potential applications involving monitoring over an extended time, it's worth noting that we also observe some systematic changes in the distribution of cognitive abilities over repeat play (Fig SI.59-68). This drift is partially due to learning effects: the games Electron Rush, Shadow Match, and Chemical Chaos show strategy heterogeneity effects at later playthroughs, as demonstrated by the high rate of extreme estimates at higher playthroughs for constructs that depend heavily on indicators from these games. Here we excluded observations that appear to be from distinct minority play strategies in these games, deferring detailed study of strategy heterogeneity and learning effects to future work. We are planning a more stringent test-retest set-up, in which we control the time between playthroughs to neutralize learning effects and ensure all the games have been played in both playthroughs.This would provide a more properreliability measurethat would befit to evaluate againstthe standard. Currently, we have only trained models on the first playthrough. It is not unreasonable to expect that we could achieve even more consistent estimates by training models dependent on the playthrough number, compensating for learning effects due to the player familiarizing themselves with the tasks.

As an example of what our Skill Lab models can already allow us to do, we used our population sample to replicate previous findings regarding the age distribution of cognitive abilities by age. Our study offers a cross-sectional snapshot of the Danish population, comprising the largest open normative dataset of these cognitive abilities. The observed patterns (Fig 7) follow the previously established expectations (Lindenberger, 2014; Salthouse, 2019), which supports Skill Lab's validity as an assessment tool. This dataset may serve as a normative benchmark for future applications, not only within psychology but also for the social sciences, clinical applications, and education. These finely stratified age norms will be of particular importance when Skill Lab addresses questions that require age-based controls.

If we want to establish more general age norms than those we have collected on the Danish population, we would naturally have to expand our recruitment efforts. As part of these efforts we have prepared a Spanish translation of Skill Lab in addition to the Danish and English translations that already existed. Furthermore, we have worked on improving the support structure around the games. This entailed improving the instructions introducing the games and removing the narrative. The narrative introduced a meta-game with a detective story that in a branching narrative guided the players from game to game. User feedback, however, told us that it hindered progress by being too hard, confusing, and buggy. With these improvements implemented we are currently planning to launch the game internationally.

Parallel to our efforts of scaling the recruitment we are also working on improving our modelling of the cognitive abilities as measured by the games. An alternative to the approach we present in this paper of aggregating indicators from multiple tasks is testing the feasibility of predicting individual task indicators from game data, which is more in line with the conventional literature (Salthouse, 2011). However, predicting individual indicators is not very robust, so we made the pragmatic choice of defining aggregated cognitive abilities measures

(Bollen & Bauldry, 2011) while being careful only to combine task indicators associated with a cognitive ability in the literature to strengthen its interpretation. In the current work, we exposed these choices to potential disconfirmation by examining their agreement across independent estimates and reject four of fourteen while accepting ten: since the data set is open, it is also open for a preliminary exploration of alternative choices. We are taking preliminary steps in this direction by pursuing a theory-driven approach, in which we only include the game indicators that are theoretically associated with a specific cognitive ability during the fitting process. The results are qualitatively similar to the ones presented here but somewhat lower in quantitative effects as expected from a restricted model. Further work in this direction may help the iterative development toward games that are optimally suited for high-quality assessment of each ability.

The models that have already been developed through our work with Skill Lab has illustrated the viability of a crowdsourcing approach in validating a cognitive assessment tool, which has several key implications. First, it allows scientists to create better human cognition models and test and validate cognitive abilities, potentially providing efficient ways to scale insights into particular cognitive abilities and how they are related to solving problems (Woolley et al., 2010). Second, we have generated a unique and open dataset, which includes normative benchmarks, that can be used as a basis for other studies. Finally, Skill Lab allows normative data for diverse populations, cultures, and languages to be collected in the future, facilitating the much-needed broadening of the samples typically tested in psychological and social science studies (Henrich et al., 2010). An advantage of Skill Lab over the traditional tests is that it is faster to play all six games once than to go through all the traditional cognitive tasks. Thus, the games could provide a low-cost self-administered test suitable for extensive deployment. This could be of great value to, e.g., the psychiatric sector in which current cognitive test batteries are burdensome to administer (Baune et al., 2018), leading to cognitive impairments often going unrecognized (Groves et al., 2018; Jaeger et al., 2006).

Acknowledgments
The authors acknowledge funding from the ERC, H2020 grant 639560 (MECTRL) and the Templeton, Synakos, Novo Nordic and Carlsberg Foundations. We would like to thank the Danish Broadcast Company DR for their collaboration without which the recruitment to the study would not have been as successful as it was. We would also like to thank the ScienceAtHome team and developers for making their contribution in designing and developing Skill Lab: Science Detective. Furthermore, we would like to acknowledge Susannah Goss for her help with copy editing, Michael Bang Petersen for commenting on the results, and Steven Langsford for his comments and help in the editing of the manuscript.

References

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

Bauer, P. J. (2020). Expanding the reach of psychological science. *Psychological Science*, *31*(1), 3–5. https://doi.org/10.1177/0956797619898664

Baune, B. T., Malhi, G. S., Morris, G., Outhred, T., Hamilton, A., Das, P., Bassett, D., Berk, M., Boyce, P., Lyndon, B., Mulder, R., Parker, G., & Singh, A. B. (2018). Cognition in depression: Can we THINC-it better? *Journal of Affective Disorders*, *225*, 559–562. https://doi.org/10.1016/j.jad.2017.08.080

Beaujean, A. A., & Benson, N. F. (2019). Theoretically-Consistent Cognitive Ability Test Development and Score Interpretation. *Contemporary School Psychology*, *23*(2), 126–137. https://doi.org/10.1007/s40688-018-0182-1

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, *55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. PubMed. https://doi.org/10.1037/a0024448

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. https://doi.org/10.1371/journal.pone.0100662

Brown, L. A., Forbes, D., & McConnell, J. (2006). Limiting the use of verbal coding in the visual patterns test. *Quarterly Journal of Experimental Psychology*, *59*(7), 1169–1176. https://doi.org/10.1080/17470210600665954

Burman, P. (1989). A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503–514. https://doi.org/10.1093/biomet/76.3.503

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Carroll, J. B. (1993). Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press.

Coughlan, G., Coutrot, A., Khondoker, M., Minihane, A.-M., Spiers, H., & Hornberger, M. (2019). Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proceedings of the National Academy of Sciences*, *116*(19), 9285–9292. https://doi.org/10.1073/pnas.1901600116

Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V. D., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2018). Global determinants of navigation ability. *Current Biology*, *28*(17), 2861-2866.e4. https://doi.org/10.1016/j.cub.2018.06.009

Danmarks nye superhjerne—DR Retrieved: 2020-07-07 https://www.dr.dk/nyheder/viden/nysgerrig/tema/danmarks-nye-superhjerne. (2020). https://www.dr.dk/nyheder/viden/nysgerrig/tema/danmarks-nye-superhjerne

Danmarks Statistik. (2018). Adgang til internettet. In *It-anvendelse i befolkningen—2018* (p. page 17). https://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=29448&sid=itbef2018

Danmarks Statistik. (2020). *Befolkningspyramide, http://extranet.dst.dk/pyramide/pyramide.htm#!y=2018&v=2* [Data set]. http://extranet.dst.dk/pyramide/pyramide.htm#!y=2018&v=2

Deary, I. J. (2011). Intelligence. *Annual Review of Psychology*, *63*(1), 453–482. https://doi.org/10.1146/annurev-psych-120710-100353

Deary, I. J., Liewald, D., & Nissan, J. (2011). A free, easy-to-use, computer-based simple and four-choice reaction time programme: The Deary-Liewald reaction time task. *Behavior Research Methods*, *43*(1), 258–268. https://doi.org/10.3758/s13428-010-0024-1

Drucker, S. M., Fisher, D., Sadana, R., Herron, J., & schraefel, m. c. (2013). TouchViz: A case study comparing two interfaces for data analytics on tablets. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2301–2310. https://doi.org/10.1145/2470654.2481318

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*(2–3), 101–118. https://doi.org/10.1080/13506289508401726

Fellows, R. P., Dahmen, J., Cook, D., & Schmitter-Edgecombe, M. (2017). Multicomponent analysis of a digital trail making test. *The Clinical Neuropsychologist*, *31*(1), 154–167. https://doi.org/10.1080/13854046.2016.1238510

Ganis, G., & Kievit, R. (2015). A new set of three-dimensional shapes for investigating mental rotation processes: Validation data and stimulus set. *Journal of Open Psychology Data*, *3*. https://doi.org/10.5334/jopd.ai

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Groves, S. J., Douglas, K. M., & Porter, R. J. (2018). A Systematic Review of Cognitive Predictors of Treatment Outcome in Major Depression. *Frontiers in Psychiatry*, *9*. https://doi.org/10.3389/fpsyt.2018.00382

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science.

Hagler, S., Jimison, H. B., & Pavel, M. (2014). Assessing executive function using a computer game: Computational modeling of cognitive processes. *IEEE Journal of Biomedical and Health Informatics*, *18*(4), 1442–1452. https://doi.org/10.1109/JBHI.2014.2299793

Hawkins, G. E., Rae, B., Nesbitt, K. V., & Brown, S. D. (2013). Gamelike features might not improve data. *Behavior Research Methods*, *45*(2), 301–318. https://doi.org/10.3758/s13428-012-0264-3

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hunt, L. T., Rutledge, R. B., Malalasekera, W. M. N., Kennerley, S. W., & Dolan, R. J. (2016). Approach-induced biases in human information sampling. *PLOS Biology*, *14*(11), e2000638. https://doi.org/10.1371/journal.pbio.2000638

Jaeger, J., Berns, S., Uzelac, S., & Davis-Conway, S. (2006). Neurocognitive deficits and disability in major depressive disorder. *Psychiatry Research*, *145*(1), 39–48. https://doi.org/10.1016/j.psychres.2005.11.011

Jennett, C., Furniss, D. J., Iacovides, I., Wiseman, S., Gould, S. J. J., & Cox, A. L. (2014). Exploring citizen psych-science and the motivations of errordiary volunteers. *Human Computation*, *1*(2). https://doi.org/10.15346/hc.v1i2.10

Jensen, A. R. (1998). The g factor: The science of mental ability. Praeger.

Kaller, C., Unterrainer, J., Kaiser, S., Weisbrod, M., Aschenbrenner, S., & Debelak, R. (2011). *Manual. Tower of London—Freiburg Version.* Vienna test system.

Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., & de Haan, E. H. F. (2000). The corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, *7*(4), 252–258. https://doi.org/10.1207/S15324826AN0704_8

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2017). *Behavioral genetics* (Seventh edition). Worth Publishers, Macmillan Learning.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Leduc-McNiven, K., White, B., Zheng, H., D McLeod, R., & R Friesen, M. (2018). Serious games to assess mild cognitive impairment: 'The game is the assessment.' *Research and Review Insights*, *2*(1). https://doi.org/10.15761/RRI.1000128

Lee, H.-J., Yost, B. P., & Telch, M. J. (2009). Differential performance on the go/no-go task as a function of the autogenous-reactive taxonomy of obsessions: Findings from a non-treatment seeking sample. *Behaviour Research and Therapy*, *47*(4), 294–300. https://doi.org/10.1016/j.brat.2009.01.002

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment*. Oxford University Press.

Lieberoth, A., Pedersen, M. K., Marin, A. C., & Sherson, J. F. (2014). Getting humans to do quantum optimization—User acquisition, engagement and early results from the citizen cyberscience game Quantum Moves. *Human Computation*, *1*(2). https://doi.org/10.15346/hc.v1i2.11

Lindenberger, U. (2014). Human cognitive aging: Corriger la fortune? *Science*, *346*(6209), 572–578. https://doi.org/10.1126/science.1254403

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. https://doi.org/10.2196/games.5888

Lumsden, J., Skinner, A., Woods, A. T., Lawrence, N. S., & Munafò, M. (2016). The effects of gamelike features and test location on cognitive test performance and participant enjoyment. *PeerJ*, *4*, e2184. https://doi.org/10.7717/peerj.2184

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford University Press.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

McNab, F., & Dolan, R. J. (2014). Dissociating distractor-filtering at encoding and during maintenance. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 960–967. https://doi.org/10.1037/a0036013

McNab, F., Zeidman, P., Rutledge, R. B., Smittenaar, P., Brown, H. R., Adams, R. A., & Dolan, R. J. (2015). Age-related changes in working memory and the ability to ignore distraction. *Proceedings of the National Academy of Sciences*, *112*(20), 6515–6518. https://doi.org/10.1073/pnas.1504162112

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A BRIEF INTRODUCTION TO EVIDENCE-CENTERED DESIGN. *ETS Research Report Series*, *2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Muender, T., Gulani, S. A., Westendorf, L., Verish, C., Malaka, R., Shaer, O., & Cooper, S. (2019). Comparison of mouse and multi-touch for protein structure manipulation in a citizen science game interface. *Journal of Science Communication*, *18*(1), A05. https://doi.org/10.22323/2.18010205

Newell, A. (1966). *On the Analysis of Human Problem Solving Protocols* [Microform]. Distributed by ERIC Clearinghouse.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Papp, K. V., Snyder, P. J., Maruff, P., Bartkowiak, J., & Pietrzak, R. H. (2011). Detecting subtle changes in visuospatial executive function and learning in the amnestic variant of mild cognitive impairment. *PLoS ONE*, *6*(7), e21688. https://doi.org/10.1371/journal.pone.0021688

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *20*(1), 33–65. https://doi.org/10.1016/j.acn.2004.02.005

Radford, J., Pilny, A., Reichelmann, A., Keegan, B., Welles, B. F., Hoye, J., Ognyanova, K., Meleis, W., & Lazer, D. (2016). Volunteer science: An online laboratory for experiments in social psychology. *Social Psychology Quarterly*, *79*(4), 376–396. https://doi.org/10.1177/0190272516675866

Reinecke, K., & Gajos, K. Z. (2015). Labinthewild: Conducting large-scale online experiments with uncompensated samples. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 1364–1378. https://doi.org/10.1145/2675133.2675246

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Rutledge, Robb B., Smittenaar, P., Zeidman, P., Brown, H. R., Adams, R. A., Lindenberger, U., Dayan, P., & Dolan, R. J. (2016). Risk taking for potential reward decreases across the lifespan. *Current Biology*, *26*(12), 1634–1639. https://doi.org/10.1016/j.cub.2016.05.017

Sagarra, O., Gutiérrez-Roig, M., Bonhoure, I., & Perelló, J. (2016). Citizen science practices for computational social science research: The conceptualization of pop-up experiments. *Frontiers in Physics*, *3*. https://doi.org/10.3389/fphy.2015.00093

Salthouse, T. A. (2011). What cognitive abilities are involved in trail-making performance? *Intelligence*, *39*(4), 222–232.

Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and Aging*, *34*(1), 17–24. https://doi.org/10.1037/pag0000288

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

Smittenaar, P., Rutledge, R. B., Zeidman, P., Adams, R. A., Brown, H., Lewis, G., & Dolan, R. J. (2015). Proactive and reactive response inhibition across the lifespan. *PLOS ONE*, *10*(10), e0140383. https://doi.org/10.1371/journal.pone.0140383

Teki, S., Kumar, S., & Griffiths, T. D. (2016). Large-scale analysis of auditory segregation behavior crowdsourced via a smartphone app. *PLOS ONE*, *11*(4), e0153916. https://doi.org/10.1371/journal.pone.0153916

Treisman, A. (1977). Focused attention in the perception and retrieval of multidimensional stimuli. *Perception & Psychophysics*, *22*(1), 1–11. https://doi.org/10.3758/BF03206074

Turkyılmaz, M. D., & Belgin, E. (2012). Reliability, Validity, and Adaptation of Computerized Revised Token Test in Normal Subjects. *Journal of International Advanced Otology*, *8*, 103–112.

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

Verbruggen, F., & Logan, G. D. (2008). Automatic and controlled response inhibition: Associative learning in the go/no-go and stop-signal paradigms. *Journal of Experimental Psychology: General*, *137*(4), 649–672. https://doi.org/10.1037/a0013170

Watson, D., Hancock, M., Mandryk, R. L., & Birk, M. (2013). Deconstructing the touch experience. *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces*, 199–208. https://doi.org/10.1145/2512349.2512819

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zysset, S., Müller, K., Lohmann, G., & von Cramon, D. Y. (2001). Color-word matching stroop task: Separating interference and response conflict. *NeuroImage*, *13*(1), 29–36. https://doi.org/10.1006/nimg.2000.0665