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], Kim Klyver[9,10], Nicolai Ladegaard[11], Kim Mathiasen[11], Christine Parsons[7], Michael Bang Petersen[12], Janet Rafner[1], Anders Ryom Villadsen[13], Mikkel Wallentin[14], Jacob Friis Sherson*[1], and Skill Lab players

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 Entrepreneurship & Relationship Management, University of Southern Denmark, Kolding, Denmark. 10. Entrepreneurship, Commercialization and Innovation Centre (ECIC), University of Adelaide, Adelaide, Australia. 11. Department of Clinical Medicine – Department of Affective Disorders, Aarhus University Hospital, Aarhus, Denmark. 12. Department of Political Science, Aarhus University, 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@phys.au.dk

Summary Paragraph

Psychology and the social sciences are undergoing a revolution: It has become increasingly clear that traditional lab-based experiments fail to capture the full range of differences in cognitive abilities and behaviours 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 employed in practical, clinical, and research contexts. Here, we present Skill Lab, a game-based approach allowing the efficient assessment of a suite of cognitive abilities. Skill Lab has been validated outside the lab in a crowdsourced population-size 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 population sample. Furthermore, by combining the game data with an in-game survey, we demonstrate that this unique dataset has implication for key questions in social science, challenging the Jack-of-all-Trades theory of entrepreneurship and provide evidence for risk preference being independent of executive functioning.

Introduction

Individual cognitive phenotyping holds the potential to revolutionize various domains ranging from personalized learning to precision psychology and the job market of the 21st century. To get there it will require us to rethink how we study and measure cognitive abilities. Most of what social scientists know about cognitive abilities and psychological behaviour has been gleaned from studying university undergraduates in the laboratory. It has become clear, however, that many of these – often underpowered – results may not generalize across populations, let alone to samples from non-Western cultures. The social sciences are therefore undergoing a revolution to increase the diversity of those studied[1]. Furthermore, in-person testing is costly, inconvenient to participants, and sometimes confounded by issues such as experimenter expectations and behaviours[2]. Until these problems are solved, individual cognitive phenotyping and ambitions such as precision psychiatry will remain an illusion.

Online crowdsourcing has been proposed as a solution to these challenges; to date, development efforts have centred around two distinct paradigms: digital versions of traditional tasks and game-based assessment. Projects such as LabintheWild[3], Volunteer Science[4], and TestMyBrain[5] offer a broad suite of digitized tasks from cognitive and social science; researchers create and post their tasks online, to be completed by volunteers from the general public. These scientific platforms have proven immensely successful for crowdsourcing data from standardized and quickly customizable tasks as an alternative to both laboratory studies and generic crowdsourcing platforms such as Amazon Mechanical Turk (MTurk). Spurred on by this success, researchers and students alike have increasingly begun to use fee-based online services to conduct studies in the social and cognitive sciences. However, many tasks, which elicit reliable within-participant effects, may actually evoke too little variation between participants to offer reliable phenotyping[6]. This is largely because, in many cases, the available tasks rely on less-stimulating and generally time-consuming and repetitive conditions, in stark contrast to the reality of our daily lives.

A broad spectrum of research indicates that games, when following evidence-centred design[7], can offer as much information about cognitive abilities as laboratory tasks designed solely for that purpose, while engaging larger and more diverse participant pools[8,9]. Prominent examples are Sea Hero Quest[10] and The Great Brain Experiment[11]. These projects motivate players by framing the game as an entertaining method to contribute to a meaningful scientific question[12,13]. Sea Hero Quest has reached 2.5 million participants and yielded important insights into spatial navigation impairments in adults at risk of Alzheimer's disease[14]; The Great Brain Experiment has provided new insights into age-related changes in working memory performance[15] and patterns of bias in information-seeking behavior[16]. These studies have demonstrated the viability of large-scale cognitive ability testing[11], but have relied on small, laboratory-based validation samples of their gamified cognitive ability measures. This raises an important question: Can we advance crowdsourced psychological science by motivating large groups of players to both play the games and perform the less entertaining and more time-consuming traditional tasks in order to provide a robust within-subject validation of game-based cognitive ability measures?

Here, we develop the most comprehensive crowdsourced validation set of cognitive ability measures to date; it could be used as a cost-effective screening tool for clinical disorders and applied in educational and corporate settings. Our broad 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. Methodologically, we address an important gap: we perform the first large-scale validation of gamified cognitive ability measures for individuals completing both digitized traditional measures and our game versions. Crowdsourcing of participant samples using MTurk has inherent challenges[17], we demonstrate that engagement of the broader "volunteer" population is possible with a rigorously designed set of gamified cognitive ability measures.

Having successfully validated our measures on a large scale, we establish a new paradigm in which the database of game-based cognitive profiles is compared with survey-based responses to a number of fundamental social science questions *by the same participant*s. Furthermore, our validation process represents a clear advance for the field of psychological science, as we move both validation and population-scale assessment outside the lab. Skill Lab is unique among big data initiatives, as it openly asks participants to contribute to scientific knowledge creation while providing them with a personalized cognitive profile based on their game play[18]. Thus, in contrast to most social science experiments, where participants' main benefit is monetary compensation, Skill Lab players' efforts are rewarded by personal feedback and an enjoyable experience[12]. Finally, as a first among big data projects in cognitive science[3–5,10,11,14–16,19–23], an anonymised version of the dataset will be openly accessible.

Skill Lab: Science Detective

Skill Lab: Science Detective is a portfolio of six games and 14 validated cognitive ability tasks (see Supplementary Information). Whereas many traditional cognitive ability tasks assess a single ability under strict conditions that aim to minimize distractions and maximize experimental control, the Skill Lab games are designed to engage multiple cognitive processes in more realistic contexts, simultaneously measuring multiple abilities within a convenient, engaging, and scalable package. We

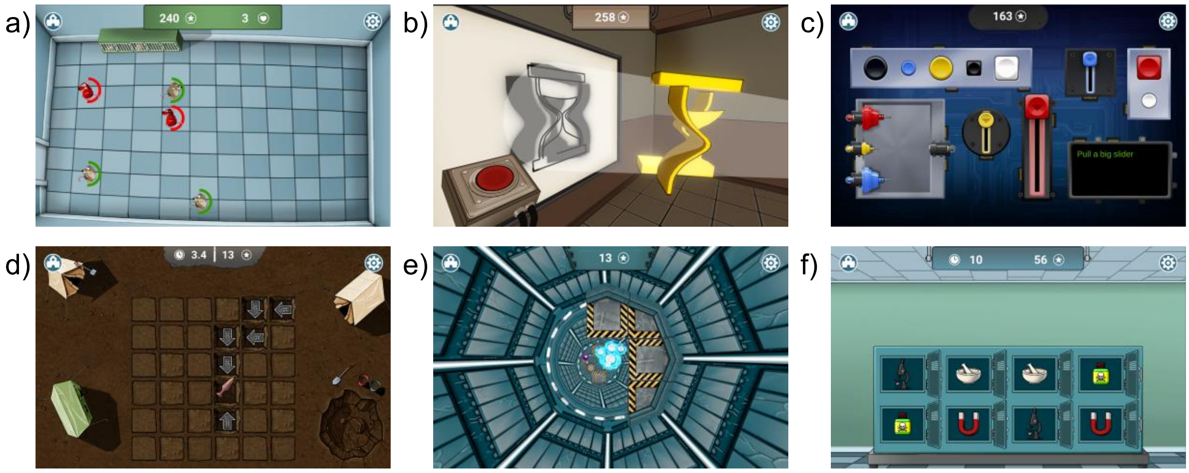

Fig 1: The six games making up Skill Lab. a) Rat Catch is designed to test Response Inhibition, Baseline Reaction Time, and Choice Reaction Time, b) Shadow Match to test visuospatial reasoning in 3D, c) Robot Reboot to test reading comprehension and instruction following, d) Relic Hunt to test visuospatial 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.

first identified how cognitive abilities have been operationalized and measured in laboratories[24–36] and designed games around specific indicators of 14 different cognitive abilities (see Supplementary Information). To ensure the validity of the cognitive abilities measured via the six games (Fig 1a–f), we administered 14 standard cognitive ability tasks in a separate section of Skill Lab. To obtain quantifiable measures of a player's level of ability, we identified *indicators* of the cognitive abilities assessed (e.g., number of errors in a task) in both the games and the tasks.

Let us illustrate the gamification process — making games out of tasks — by describing the relationship between the classic Go/No-Go task[29] and the Rat Catch game (Fig. 1b; see Supplementary Information for descriptions of the other games). The Go/No-Go task 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 colour, then the participant must react as quickly as possible; otherwise, the participant must refrain from reacting. This test procedure is mirrored 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 must tap the rat as quickly as possible, providing measures of Baseline Reaction Time. The rats disappear faster and faster as the level progresses; when the player has missed three rats, they are sent to the next level.

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 provides indicators of Choice Reaction Time and Response Inhibition. Further levels of Rat Catch add variations, such as an increasing number of stimuli or moving targets. These additions give indicators of visuospatial reasoning components, such as 2D spatial representation and movement perception. Through the scripted behavioural pattern assessment[7] 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).

Validating cognitive abilities "in the wild"

We have two separate participant samples for Skill Lab: i) an initial sample recruited through MTurk (n = 444) and ii) more than 18,000 people who signed up to play the publicly available version (Fig. 2a). Having both groups enables us to demonstrate the challenges and benefits of crowdsourced validation

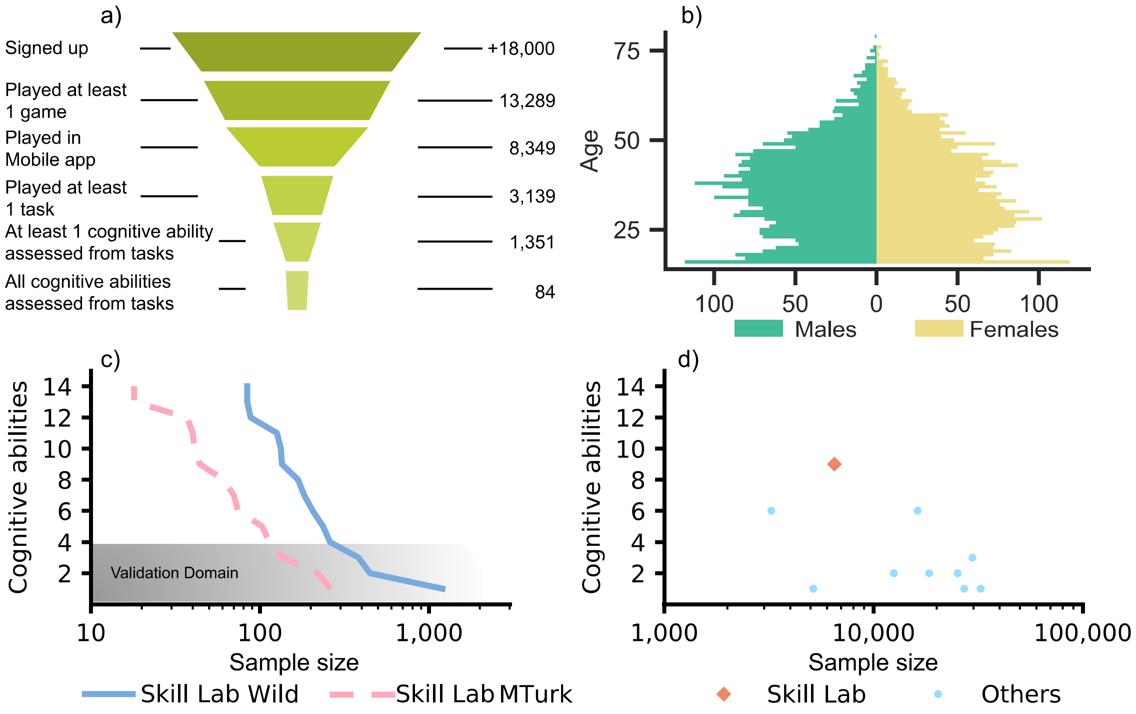

Fig 2: a) Funnel of wild player recruitment. At each layer of the funnel, fewer players had chosen to play. A small minority of players reached the bottom layer, providing enough data for us to assess all cognitive abilities from the tasks. b) Age and gender distribution for players who played at least 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[24–36]. d) Sample size and number of cognitive abilities measured: Skill Lab games compared with other population-scale assessment studies[10,11,14–16,19–23]

in the wild. The MTurk study was split into six separate jobs — one for each game and the associated tasks. We recruited 100 MTurkers per job — they were allowed to participate in multiple different jobs — over a 2-week period at the end of June 2018; our sample was thus about 3–4 times larger than that used for within-subject validation in Sea Hero Quest[14] or The Great Brain Experiment[19].

MTurk's terms of service only allow data collection via an in-browser version of Skill Lab. Thus, an app version of Skill Lab could only be validated in the wild. Because participant engagement typically has an exponential fall off[37], and because we needed players to both play the games and complete the validation tasks, we sought to recruit as many people as possible. Skill Lab launched publicly in Denmark in collaboration with the Danish Broadcast Company (Danmarks Radio, DR), 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[38]; thus, to attract the broadest possible audience, we generated attention to the project through a series of DR news articles with themes varying from AI and technology to psychology and computer games[39]. Participants who played at least one game represent a broad cross-section of the Danish population[40] in terms of gender (5793 female, 7333 male, and 163 other; or 44%, 55%, and 1%, respectively) and age (Fig. 2b), 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; of those, 38% completed at least one cognitive ability task (Fig. 2a). 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 in order for us to evaluate the ability Visuospatial Working Memory). Even with these requirements, we obtained a larger sample of wild players for the cognitive ability measures than from MTurk (Fig. 2c). MTurk participants often sacrifice accuracy for speedy task completion. We found that this was not the case for wild players (see Supplementary Information). The games were specifically developed to motivate players to do their best, and it is faster to complete all the games combined (14 ± 5 min) than all the cognitive ability tasks combined (72 ± 7 min). Thus, the games could potentially be used for rapid cognitive assessments.

To validate the cognitive ability measures from the games trained a linear model to predict cognitive abilities - as measured by the tasks - from game indicators. To obtain estimates of the out-of-sample prediction strengths, we applied repeated cross-validation[41] and used an elastic-net to avoid overfitting by performing variable selection and mitigating multicollinearity[42]. The process resulted in nine accepted prediction models with medium to strong effect sizes (Fig. 3a) and five rejected models; four were rejected because the model collapsed to the mean value (Fig. 3c), and one because of ceiling and discretisation effects of the task measure. Although it is possible that more advanced modelling of the existing data set can improve these results, the nine accepted models already represent a broad, strong, and rapid testing battery, ready for application.

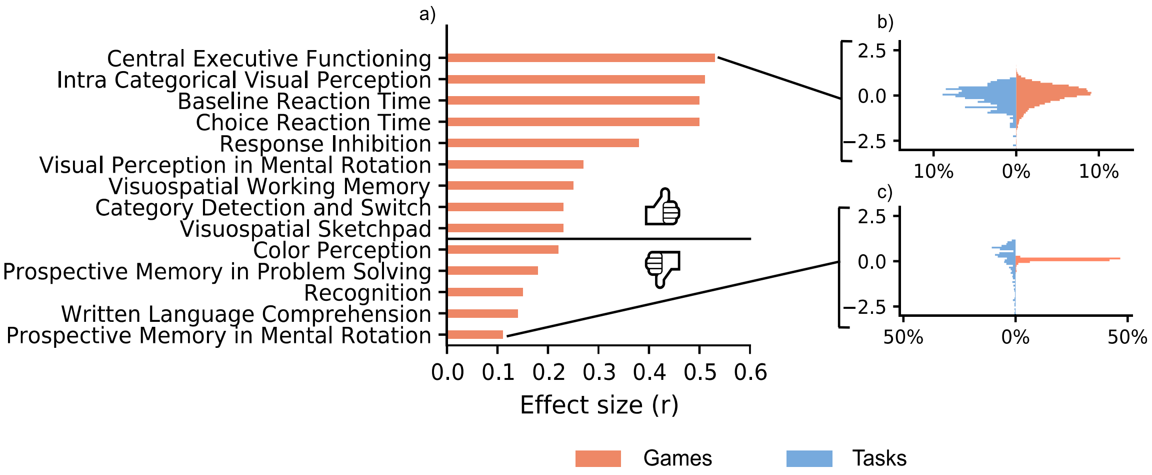

Fig 3: a) Out-of-sample correlation strength for the elastic-net models predicting task-based cognitive ability measures from Skill Lab game indicators. The nine models above the black line were accepted. b) Population distribution of Central Executive Functioning from the tasks (blue) and the games (orange). X-axis: percentage of Skill Lab wild players with a specific cognitive ability measure level as measured from the tasks/games. c) Population distribution of Prospective Memory in Mental Rotation from the tasks (blue) and the games (orange). X-axis: percentage of Skill Lab wild players with a specific cognitive ability measure level as measured from the tasks/games.

Population-scale assessment

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[10,11,14–16,19–23]; 6,312 players had played enough for the nine trained models to be applied. We have primarily collected data within the Danish population, but Skill Lab is ready for follow-up studies in other countries as it is available in Danish, English, and Spanish. It approaches the population-scale assessment usually limited to registry, commercial, and meta-studies, thereby providing a unique tool.

Our population sample allows us to both replicate previous studies and address new questions in cognitive and social science. Most cognitive abilities increase during childhood and adolescence and then begin to decline in the mid-20s to 30s[43,44], with a few cognitive abilities such as verbal fluency remaining constant through adulthood. Skill Lab provides two distributions across age for each cognitive ability—one measured by the tasks and one measured by the games (Fig. 4). Given our pattern of participation, the sample size for the games (n = 6,312) was significantly larger than that for the tasks (on average n = 311), providing considerably more data to resolve trends and remove noise.

Our study offers a cross-sectional snapshot of the Danish population, comprising the largest open normative dataset of these cognitive abilities. 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, which provides further support for the validity of Skill Lab 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 is used to address questions that require age-based controls. In addition, we can extract key indicators such as age of onset of decline and crossing of the general population average, which can then be applied directly in clinical and other settings. We use the age-stratified norms of the cognitive abilities to control for age effects in all the following analyses.

By linking cognitive profiles with survey data obtained from the same participants on entrepreneurship and risk preferences (see Supplementary Information), we were able to generate a unique dataset with the potential to generate new knowledge in social science.

The survey included a question regarding players' entrepreneurial intention—that is, the degree to which they were interested in starting their own business. It has been argued that entrepreneurial intention should be negatively correlated with cognitive abilities due to the opportunity costs of alternative employment options[45,46]; that is, people with high cognitive abilities are more likely to have good employment opportunities. Our findings confirmed this hypothesis ($r = -0.09$, $p = 0.01$, $n = 720$), with all individual abilities showing a negative correlation with entrepreneurial intention, and most correlations being significant. Surprisingly, contrary to the Jack-of-all-Trades theory[45–47], which predicts that generalists (who have a uniform distribution of abilities) have a better fit with an entrepreneurial career, we found that people with greater variation in cognitive abilities had higher entrepreneurial intention ($r = 0.09$, $p = 0.01$, $n = 720$). Prior tentative confirmations of the Jack-of-all-Trades theory have

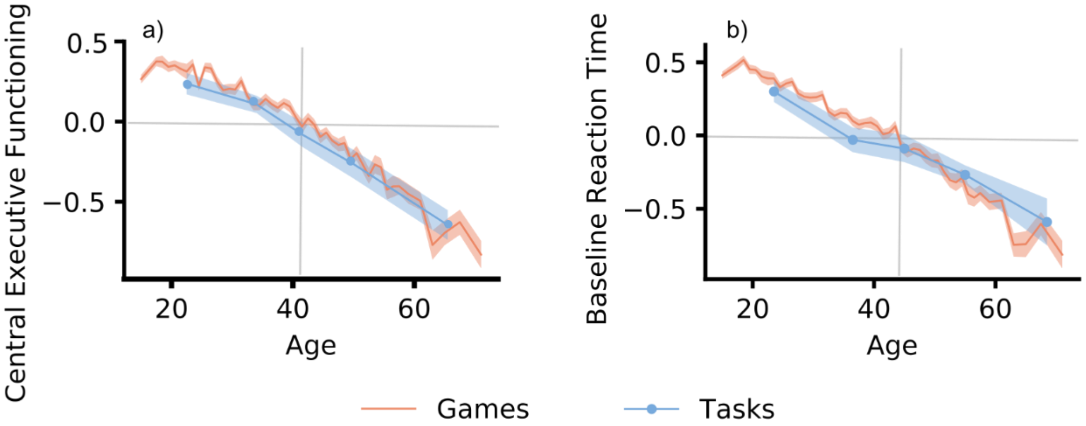

Fig 4: Cognitive abilities across age groups. 6,312 wild players played the games; fewer played the combination of tasks that allowed for assessment of a specific ability. The shaded areas around the curves are the standard error of the mean. Each age point in the graph includes at least 30 players (the curves for the remaining cognitive abilities can be found in the Supplementary Information). The grey lines indicate where the population crosses zero. a) Central Executive Functioning (n_{task} = 254), b) Baseline Reaction Time (n_{task} = 228

relied on measurements of practical skills, e.g. math, logic, language, or technical abilities[43,44]. Since our findings are based on measurements of cognitive abilities, we speculate that individuals with greater variation in cognitive abilities are more likely to identify a better match to entrepreneurship, where they themselves have the autonomy to define their functions[48]. Integrating data on lower-level cognitive abilities thus challenges the Jack-of-all-trade theory's distinction between generalists vs. specialists, which expands our insights into the characteristics of entrepreneurs.

Risk preference is assumed to contribute to key life outcomes across many domains[49]. There is an ongoing debate about whether risk preference varies systematically with cognitive abilities, in particular executive function[49]. If this were the case, it would have intriguing policy implications, as individuals' cognitive abilities would need to be accounted for[49]. In previous studies, small sample size and lack of power have left this important matter unsettled. In the survey part of our study, we therefore administered three typical risk measures[50]: two behavioural and one self-reported. Correlations between the risk measures and six measures of executive function were not significant, and Bayes Factors[51] provide strong evidence — by far the strongest evidence to date — for the absence of any effect of executive function on risk preferences (average $BF_{01} = 20.0$, $n = 920$).

Our work with Skill Lab has illustrated the viability of a crowdsourcing approach in validating a cognitive assessment tool and has several key implications. First, it allows scientists to create better models of human cognition and to test and validate cognitive abilities, potentially providing insights into more efficient ways of solving problems[52]. Second, our unique and open dataset, which includes normative benchmarks, can potentially inform large-scale screening for the development of psychological disorders. 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[53].

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 Mario Alba for his help and discussions on the theoretical foundations, and Susannah Goss for her help with copy editing.

Methods

Tasks administered in Skill Lab

Corsi Block[24],Deary-Liewald[25], Eriksen-Flanker[26], Groton Maze[27]. Mental Rotation[28], Go/No-Go[29], Stop Signal[30], Stroop[31], Token Test[36], Tower of London[32], Trailmaking[33], Visual Pattern[35], Visual Search Letters[34], and Visual Search Shapes[34].

MTurk

The MTurk sample was collected by publishing six different MTurk tasks (Human Intelligence Task, or HIT). In each HIT, MTurkers played one of the games as well as the tasks assessing the cognitive ability hypothesized to be associated with that game (see Fig. SI.14). Before launching the jobs, we made a power calculation of the sample size required for Pearson correlations to measure medium effect sizes ($r > 0.3$), which showed that we needed a sample of at least 85 MTurkers. To allow for removal of some outliers, we decided to recruit 100 MTurkers per HIT. We started by publishing an initial batch of nine jobs for each of the HITs in order to determine the completion time and thus what a fair payment would be. We found that each task took approximately 40 minutes to complete and thus settled on a payment of US$6 per task (US$9/hour average); in addition, we offered a bonus for completing multiple HITs:

- A 5% bonus on the second HIT.
- A 10% bonus on the third HIT.
- A 15% bonus on the fourth HIT.
- A 20% bonus on the fifth HIT.
- A 25% bonus on the sixth HIT.

Furthermore, if MTurkers had already completed a cognitive ability task previously, they did not have to retake the task, which enabled them to complete the job faster and thus increase their hourly wage. Each MTurker was only allowed to take one job from each of the six HITs.

The tasks were released in batches of nine jobs for all HITs at the same time. The batches were released at irregular intervals at all times of day from 27th July 2018 to 2nd August 2018. The MTurkers were required to have at least 500 previously approved HITs and a 90% approval rate. We did not put any regional restrictions on the HITs.

Modeling cognitive abilities with games

We trained a model that predicts players' cognitive abilities measured from the tasks based on how they played the games by fitting a linear. For each task, multiple indicators t_i of a cognitive ability were computed from the data (see Supplementary Information). We reviewed the tasks[24–36] to identify how the *i*'th task indicator *tᵢ*, contributed to the measure of the *j*'th cognitive ability C_j by assigning a coefficient α_{ij} ∊ {-1,0,1}: 0if there is no contribution, 1 if there is a positive correlation between the task indicator and a higher level cognitive ability, and -1 if there is a negative correlation (see Supplementary Information for a comprehensive list of coefficients). The task indicators were combined into measures of cognitive abilities[54] by taking weighted (α_{ij}) averages

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

In total, 46 indicators gi from the six games were identified as containing information pertaining to the cognitive abilities. Before any modelling was performed, all game indicators and cognitive ability measures were standardized to mean = 0 and SD = 1 to put them on equal footing, and values more than 3 SD from the mean were excluded as outliers. Only players who had produced all the task indicators associated with respective cognitive ability (see Suplementary Information) as well as at least one game indicator were included in the sample used to fit the linear regression models predicting the cognitive abilities measured from the tasks with game indicators. Any missing game indicators were imputed using multivariate imputation with chained equations[55], 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. In order 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. The trained models ({β_{1j},...,β_{45j}}, k_j) (see Supplemental Information) would be the result of averaging all the 500 individually trained models per cognitive ability. We have an estimated out-of-sample prediction strength defined as the Pearson correlation between the predicted values of each of the models and the cognitive abilities from the tasks for each of the repeated-cross validation test sets (Table 1).

Distributions of cognitive abilities across age

The age data points in Fig. 4 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 from the task-measured cognitive abilities. The points were generated by starting at age 16 and checking whether there were 30 players of that age whose data provided a cognitive ability measure. If there were enough, 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.

For the age-corrected normative data used to control for age in the rest of the paper (see Supplementary Information), we defined 5-year

Cognitive Ability	n	r	95% Confidence Interval	p
Central Executive Functioning	191	0.53	[0.42, 0.62]	< 0.00001
Intra Categorical Visual Perception	868	0.51	[0.46, 0.56]	< 0.00001
Choice Reaction Time	65	0.50	[0.29, 0.66]	0.00001
Baseline Reaction Time	161	0.50	[0.37, 0.61]	< 0.00001
Response Inhibition	82	0.38	[0.18, 0.55]	0.00042
Visual Perception in Mental Rotation	327	0.27	[0.17 0.37]	< 0.00001
Visuospatial Working Memory	135	0.25	[0.08, 0.40]	0.00345
Visuospatial Sketchpad	204	0.23	[0.10, 0.36]	0.00093
Category Detection and Switch	95	0.23	[0.03, 0.41]	0.02494
Color Perception	300	0.22	[0.11, 0.33]	0.00012
Prospective Memory in Problem Solving	124	0.18	[0.00, 0.34]	0.04545
Recognition	168	0.15	[0.00, 0.29]	0.05229
Written Language Comprehension	199	0.14	[0.00, 0.27]	0.04858
Prospective Memory in Mental Rotation	320	0.11	[0.00, 0.22]	0.04930

Table 1: Results of fitting the cognitive abilities with an elastic-net model.

	r	95% Confidence Interval	p
Intra Categorical Visual Perception	-0.11	[-0.18, -0.04]	0.002551
Central Executive Functioning	-0.08	[-0.15, -0.01]	0.032999
Visual Perception in Mental Rotation	-0.07	[-0.14, 0.00]	0.05541
Baseline Reaction Time	-0.07	[-0.14, 0.01]	0.047647
Category Detection and Switch	-0.07	[-0.14, 0.00]	0.078376
Visuospatial Working Memory	-0.05	[-0.12, 0.02]	0.143245
Visuospatial Sketchpad	-0.08	[-0.15, -0.01]	0.041798
Response Inhibition	-0.12	[-0.19, -0.05]	0.001132
Choice Reaction Time	-0.06	[-0.13, 0.01]	0.098032
AVG Cognitive Ability	-0.09	[-0.16, -0.02]	0.011273
SD Cognitive Ability	0.09	[0.02, 0.16]	0.013749

Table 2: Correlation between entrepreneurial intention and age-corrected cognitive abilities (n = 720 for all correlations)

non-overlapping intervals from ages 16–100 years. Cognitive abilities were corrected by standardizing within the 5-year age bins.

Correlation between survey data and cognitive abilities

All correlations between survey data and cognitive abilities were Pearson correlations; the correlations are provided below.

To assess entrepreneurial intention, we asked people not currently in self-employment to estimate the probability they would start their own business in the next 5 years (response options: 0%, 1–20%, 21–40%, 41–60%, 61–80%, 81–99%, 100%). We correlated these data with the cognitive abilities, as well as with the average level and standard deviation of the cognitive abilities. The latter are standard measures of the Jack-of-all-Trades theory in the entrepreneurship literature.

Risk behaviour was measured by three questions:

(SOEP) Are you generally a person who is willing to take risks or do you try to avoid taking risks? (response options: Lickert scale 0-10; 0 = Not Willing to take risks, 10 = Very willing to take risks)

(Risk-Risk) Below you are presented with a choice between two lotteries, Option 1 and Option 2. The options offer different amounts of money with different probabilities. Please read the characteristics of the options carefully and indicate—assuming that this was a real choice—how strongly you would prefer Option 1 or Option 2.

Option 1: 80% chance of winning €200 and a 20% chance of winning €160.

Option 2: 80% chance of winning €300 and a 20% chance of winning €10.

(response options: Likert scale 1-9; 1 = Strongly prefer option 1, 5 = Both options are equally attractive, 9 = Very strongly prefer option 2)

(Safe-Risk) Below you are presented with a choice between two lotteries, Option 1 and Option 2. One option offers a certain monetary reward for sure, the other option offers different amounts of money with different probabilities. Please read the characteristics of the options carefully and indicate—assuming that this was a real choice—how strongly you would prefer Option 1 or Option 2.

Option 1: €192 for sure.
Option 2: 80% chance of winning €300 and a 20% chance of winning €10.

(response options: Likert scale 1-9; 1 = Strongly prefer option 1, 5 = Both options are equally attractive, 9 = Very strongly prefer option 2)

	SOEP	Risk-Risk	Safe-Risk

Central Execution	r = 0.001 [-0.06, 0.07]
	p = 0.99
	BF_{01} = 23.8
		r = 0.01 [-0.05, 0.08]
		p = 0.69
		BF_{01} = 22.2
			r = -0.01 [-0.08, 0.05]
			p = 0.67
			BF_{01} = 21.7

Visuospatial Working Memory	r = -0.02 [-0.08, 0.05]
	p = 0.61
	BF_{01} = 20.8
		r = 0.04 [-0.03, 0.10]
		p = 0.25
		BF_{01} = 12.5
			r = 0 [-0.07, 0.06]
			p = 0.88
			BF_{01} = 23.8

Baseline Reaction Time	r = -0.01 [-0.07, 0.06]
	p = 0.85
	BF_{01} = 23.3
		r = 0 [-0.06, 0.07]
		p = 0.94
		BF_{01} = 23.8
			r = -0.03 [-0.10, 0.03]
			p = 0.31
			BF_{01} = 14.5

Category Detection and Switch	r = 0 [-0.06, 0.06]
	p = 0.99
	BF_{01} = 23.8
		r = 0.01 [-0.05, 0.08]
		p = 0.75
		BF_{01} = 22.7
			r = -0.02 [-0.09, 0.04]
			p = 0.48
			BF_{01} = 18.5

Response Inhibition	r = -0.03 [-0.10, 0.03]
	p = 0.34
	BF_{01} = 15.4
		r = -0.01 [-0.08, 0.05]
		p = 0.72
		BF_{01} = 22.2
			r = -0.04 [-0.10, 0.03]
			p = 0.27
			BF_{01} = 13.2

Choice Reaction Time	r = 0 [-0.07, 0.06]
	p = 0.96
	BF_{01} = 23.8
		r = -0.02 [-0.09, 0.04]
		p = 0.54
		BF_{01} = 19.6
			r = -0.03 [-0.09, 0.04]
			p = 0.42
			BF_{01} = 17.2

Table 3: Correlation between risk behaviour and age-corrected cognitive abilities [with 95% confidence interval] (n = 920 for all correlations). BF01 is the Bayes Factor indicating evidence for the absence of a correlation.

Game Availability

Skill Lab: Science Detective is available on the Apple App Store, Google Play, and online at https://www.scienceathome.org/games/skill-lab-science-detective/play-skill-lab/

Informed consent and ethics statement

Players both in MTurk and in the wild provided informed consent before taking part in the study and before any data were recorded. They were made aware that they could, at any time, leave the study and request their data to be anonymized.

The Committee of Research Ethics for Region Midtjylland (Denmark) exempted the study from ethical oversight, and the project received ethical approval from the Institutional Review Board at Cornell University (Protocol ID: 1808008201). The study was conducted in accordance with all ethical requirements.

Author contributions

M.K.P., C.M.C.D, R.V.B, and J.F.S developed the concept. C.M.C.D identified validation tasks and designed the games. M.K.P., C.B., and J.F.S. were in charge of data collection. M.K.P. and M.G. performed the data analysis. J.F.S. supervised the project. M.H.C. procured ethical approval from Cornell University. M.K.P wrote the manuscript. M.K.P. and C.M.C.D contributed equally to the project. All authors participated in hypothesis generation, discussed the results and implications, and commented on the manuscript at all stages.

References

1. Bauer, P. J. Expanding the reach of psychological science. *Psychol. Sci.* **31**, 3–5 (2020).
2. Birnbaum, M. H. Human research and data collection via the internet. *Annu. Rev. Psychol.* **55**, 803–832 (2004).
3. Reinecke, K. & Gajos, K. Z. Labinthewild: conducting large-scale online experiments with uncompensated samples. in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15* 1364–1378 (ACM Press, 2015). doi:10.1145/2675133.2675246.
4. Radford, J. *et al.* Volunteer science: an online laboratory for experiments in social psychology. *Soc. Psychol. Q.* **79**, 376–396 (2016).
5. Germine, L. *et al.* Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* **19**, 847–857 (2012).
6. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).
7. Shute, V. J., Wang, L., Greiff, S., Zhao, W. & Moore, G. Measuring problem solving skills via stealth assessment in an engaging video game. *Comput. Hum. Behav.* **63**, 106–117 (2016).
8. Baniqued, P. L. *et al.* Selling points: What cognitive abilities are tapped by casual video games? *Acta Psychol. (Amst.)* **142**, 74–86 (2013).
9. Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D. & Munafò, M. R. Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. *JMIR Serious Games* **4**, e11 (2016).
10. Coughlan, G. *et al.* Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proc. Natl. Acad. Sci.* **116**, 9285–9292 (2019).

11. Brown, H. R. *et al.* Crowdsourcing for cognitive science – the utility of smartphones. *PLoS ONE* **9**, e100662 (2014).
12. Sagarra, O., Gutiérrez-Roig, M., Bonhoure, I. & Perelló, J. Citizen science practices for computational social science research: the conceptualization of pop-up experiments. *Front. Phys.* **3**, (2016).
13. Jennett, C. *et al.* Exploring citizen psych-science and the motivations of errordiary volunteers. *Hum. Comput.* **1**, (2014).
14. Coutrot, A. *et al.* Global determinants of navigation ability. *Curr. Biol.* **28**, 2861–2866.e4 (2018).
15. McNab, F. *et al.* Age-related changes in working memory and the ability to ignore distraction. *Proc. Natl. Acad. Sci.* **112**, 6515–6518 (2015).
16. Hunt, L. T., Rutledge, R. B., Malalasekera, W. M. N., Kennerley, S. W. & Dolan, R. J. Approach-induced biases in human information sampling. *PLOS Biol.* **14**, e2000638 (2016).
17. Stewart, N., Chandler, J. & Paolacci, G. Crowdsourcing samples in cognitive science. *Trends Cogn. Sci.* **21**, 736–748 (2017).
18. Rudnicka, A., Cox, A. L. & Gould, S. J. J. Why do you need this? : selective disclosure of data among citizen scientists. in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* 1–11 (ACM Press, 2019). doi:10.1145/3290605.3300622.
19. Rutledge, R. B., Skandali, N., Dayan, P. & Dolan, R. J. A computational and neural model of momentary subjective well-being. *Proc. Natl. Acad. Sci.* **111**, 12252–12257 (2014).
20. McNab, F. & Dolan, R. J. Dissociating distractor-filtering at encoding and during maintenance. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 960–967 (2014).
21. Smittenaar, P. *et al.* Proactive and reactive response inhibition across the lifespan. *PLOS ONE* **10**, e0140383 (2015).
22. Teki, S., Kumar, S. & Griffiths, T. D. Large-scale analysis of auditory segregation behavior crowdsourced via a smartphone app. *PLOS ONE* **11**, e0153916 (2016).
23. Rutledge, R. B. *et al.* Risk taking for potential reward decreases across the lifespan. *Curr. Biol.* **26**, 1634–1639 (2016).
24. Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J. & de Haan, E. H. F. The corsi block-tapping task: standardization and normative data. *Appl. Neuropsychol.* **7**, 252–258 (2000).
25. Deary, I. J., Liewald, D. & Nissan, J. A free, easy-to-use, computer-based simple and four-choice reaction time programme: The Deary-Liewald reaction time task. *Behav. Res. Methods* **43**, 258–268 (2011).
26. Eriksen, C. W. The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Vis. Cogn.* **2**, 101–118 (1995).
27. Papp, K. V., Snyder, P. J., Maruff, P., Bartkowiak, J. & Pietrzak, R. H. Detecting subtle changes in visuospatial executive function and learning in the amnestic variant of mild cognitive impairment. *PLoS ONE* **6**, e21688 (2011).
28. Ganis, G. & Kievit, R. A new set of three-dimensional shapes for investigating mental rotation processes: validation data and stimulus set. *J. Open Psychol. Data* **3**, (2015).
29. Lee, H.-J., Yost, B. P. & Telch, M. J. 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. *Behav. Res. Ther.* **47**, 294–300 (2009).
30. Verbruggen, F. & Logan, G. D. Automatic and controlled response inhibition: Associative learning in the go/no-go and stop-signal paradigms. *J. Exp. Psychol. Gen.* **137**, 649–672 (2008).
31. Zysset, S., Müller, K., Lohmann, G. & von Cramon, D. Y. Color-word matching stroop task: separating interference and response conflict. *NeuroImage* **13**, 29–36 (2001).
32. Kaller, C. *et al. Manual. Tower of London - Freiburg Version.* (Vienna test system., 2011).
33. Fellows, R. P., Dahmen, J., Cook, D. & Schmitter-Edgecombe, M. Multicomponent analysis of a digital trail making test. *Clin. Neuropsychol.* **31**, 154–167 (2017).
34. Treisman, A. Focused attention in the perception and retrieval of multidimensional stimuli. *Percept. Psychophys.* **22**, 1–11 (1977).
35. Brown, L. A., Forbes, D. & McConnell, J. Limiting the use of verbal coding in the visual patterns test. *Q. J. Exp. Psychol.* **59**, 1169–1176 (2006).
36. Turkyılmaz, M. D. & Belgin, E. Reliability, Validity, and Adaptation of Computerized Revised Token Test in Normal Subjects. *J. Int. Adv. Otol.* **8**, 103–112 (2012).
37. Lieberoth, A., Pedersen, M. K., Marin, A. C. & Sherson, J. F. Getting humans to do quantum optimization - user acquisition, engagement and early results from the citizen cyberscience game Quantum Moves. *Hum. Comput.* **1**, (2014).
38. Danmarks Statistik. Adgang til internettet. in *It-anvendelse i befolkningen - 2018* page 17 (2018).
39. *Danmarks nye superhjerne - DR Retrieved: 2020-07-07 https://www.dr.dk/nyheder/viden/nysgerrig/tema/danmarks-nye-superhjerne*. (2020).
40. Danmarks Statistik. Befolkningspyramide, http://extranet.dst.dk/pyramide/pyramide.htm#!y=2018&v=2. (2020).
41. Burman, P. A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503–514 (1989).
42. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
43. Lindenberger, U. Human cognitive aging: Corriger la fortune? *Science* **346**, 572–578 (2014).
44. Salthouse, T. A. Trajectories of normal cognitive aging. *Psychol. Aging* **34**, 17–24 (2019).
45. Hartog, J., Van Praag, M. & Van Der Sluis, J. If you are so smart, why aren't you an entrepreneur? Returns to cognitive and social ability: entrepreneurs versus employees. *J. Econ. Manag. Strategy* **19**, 947–989 (2010).
46. Aldén, L., Hammarstedt, M. & Neuman, E. All about balance? A test of the Jack-of-all-Trades theory using military enlistment data. *Labour Econ.* **49**, 1–13 (2017).
47. Lazear, E. P. Entrepreneurship. *J. Labor Econ.* **23**, 649–680 (2005).

48. Holland, J. L. *Making vocational choices: a theory of vocational personalities and work environments*. (Psychological Assessment Resources, 1997).
49. Dohmen, T., Falk, A., Huffman, D. & Sunde, U. On the relationship between cognitive ability and risk preference. *J. Econ. Perspect.* **32**, 115–134 (2018).
50. Mata, R., Frey, R., Richter, D., Schupp, J. & Hertwig, R. Risk preference: a view from psychology. *J. Econ. Perspect.* **32**, 155–172 (2018).
51. Ly, A., Verhagen, J. & Wagenmakers, E.-J. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *J. Math. Psychol.* **72**, 19–32 (2016).
52. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
53. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
54. Bollen, K. A. & Bauldry, S. Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychol. Methods* **16**, 265–284 (2011).
55. Buuren, S. van & Groothuis-Oudshoorn, K. mice : multivariate imputation by chained equations in r. *J. Stat. Softw.* **45**, (2011).