OU Brainwave: a mobile app delivering a gamified battery of cognitive tests designed for repeated play

THIRKETTLE, Martin <http://orcid.org/0000-0002-6200-3130>, LEWIS, Jen <http://orcid.org/0000-0002-3765-1566>, LANGDRIDGE, Darren <http://orcid.org/0000-0001-9319-9802> and PIKE, Graham <http://orcid.org/0000-0003-1013-8698>

Published version

THIRKETTLE, Martin, LEWIS, Jen, LANGDRIDGE, Darren and PIKE, Graham (2018). OU Brainwave: a mobile app delivering a gamified battery of cognitive tests designed for repeated play. JMIR Serious Games, 6 (4).

OU Brainwave: A mobile app delivering a gamified battery of cognitive tests designed for repeated play

Abstract

Background: Mobile phone and tablet apps are an increasingly common platform to collect data. A key challenge for researchers has been participant “buy-in” and participant attrition for designs requiring repeated testing.

Objective: The objective of this study was to develop and asses the utility of 1 – 2 minute versions of both classic and novel cognitive tasks within a user focussed and driven mobile phone and tablet app designed to encourage repeated play.

Methods: A large sample (N = 13,979 at first data collection) participated in multiple, self-paced, sessions of classic working memory (N-back), spatial cognition (Mental rotation), sustained attentional focus (Persistent Vigilance task), and split attention (Multiple object tracking) tasks along with an implementation of a, comparatively, novel action learning task. The app, "OU Brainwave" was designed to measure time-of-day variation in cognitive performance, and did not offer any training program or promise any cognitive enhancement. To record participant's chronotype a full Morningness-Eveningness questionnaire was also included. Data was collected across an 18 month period. While the app prompted reengagement at set intervals, each participant was free to repeatedly complete each task as many times as they wished.

Results: We found a significant relationship between Morningness and age (r = 0.298, n = 12755, p < 0.001), though no effect of gender (t (13539) = -1.036, p = 0.30). We report good task adherence, with ~4000 participants repeatedly playing each game more than four times each - our minimum engagement level for analysis. The repeated plays of these games allow us to replicate commonly reported gender effects in the gamified spatial cognition (F (1, 4216) = 154.861, p<0.001, η_ρ^2 = 0.035), split attention (F (1, 4185) = 11.047, p=0.001, η_ρ^2 = 0.003), and sustained attentional focus (F (1, 4238) = 15.993, p<0.001, η_ρ^2 = 0.004) tasks. We also report evidence of a small gender effect in an action learning task (F (1, 3988) = 90.59, p<0.001, η_ρ^2 = 0.022). Finally, we also found strong negative correlations between self-reported age and performance in the sustained attentional focus (N=1596, F (6, 1595) = 30.23, p<0.001, η^2 = 0.102), working memory (N = 1627, F (6, 1626) = 19.78, p<0.001,η^2 = 0.068), spatial cognition (N = 1640, F (6, 1639) = 23.74, p<0.001,η^2 = 0.080)), and split attention (N = 1616, F(6,1615) = 2.48, p= 0.022, η^2 = 0.009) tasks.

Conclusions: Using extremely short testing periods and permitting participants to decide their own level of engagement - both in terms of which gamified task they played, and how many sessions they completed - we were able to collect a substantial and valid dataset. We suggest that the success of OU brainwave should inform future research oriented apps - particularly in issues around balancing participant engagement with data fidelity.
Keywords: mobile apps; cognitive psychology; Morningness-Eveningness; gamification; experimental game; behavioural research

Introduction

Recent advances in the performance and accessibility of web technologies have resulted in the increasing use of web platforms to conduct cognitive psychology research. The large and diverse cohorts easily available to researchers are now accompanied by platforms capable of implementing complex tasks and accurately measuring performance [1,2]. Moving on from interactive webpages, the possibilities offered by custom built, natively coded, mobile apps include high levels of stimulus control and enormous flexibility as regards experimental design and data collection – both in terms of frequency of data collecting sessions and the range of data collected [3]. By collecting large sets of cognitive performance data, insights into the subtle variations in cognition, both within an individual as here, or across individuals, and cultures [4] are potentially available to the researcher. Aspects of the tasks included here are prevalent in many everyday skills and activities - from attending to all the potential threats when crossing a busy road ("Track"- multiple object tracking), to efficiently packing a suitcase ("Spin" - mental rotation). Understanding cognitive performance is hugely important: even if we consider only healthy mental function, it is only by understanding the fundamental properties of our cognition, can we design our lives [5], work [6,7], and play [8] to enable our own best performance [9,10].A key issue for all psychology researchers is recruiting participants. While lab-based studies can often rely on departmental participation requirement to ensure a steady flow of, debatably, willing participants, the sample obtained is inevitably limited in terms of various demographic factors [11] . Online and app-based studies are one possible way of researching with a broader sample of participants, but to achieve this they must ensure their task, or request, is an engaging one, especially if it requires repeated testing sessions for data collection. Embedding the experimental collecting task within an engaging, fun-to-play, game is an increasingly popular way of trying to improve participant engagement and retention. A recent systematic review of gamification of cognitive tasks suggested increased engagement was one of the main reasons for gamification

[12]. Moreover, this review highlighted the additional benefits of gamification, such as reducing anxiety, and extending the reach of the researcher, whilst underlining the potential that gamification has to improve data collection without necessarily impairing data validity.

OU Brainwave is a bespoke app, launched on multiple platforms, designed to collect research data while providing participants with understandable measures of their own performance across five facets of cognitive ability. The app includes gamified tasks designed to measure performance on aspects of working memory, spatial cognition, sustained attentional focus, split attention and action learning. Importantly, we did not set out to 'train' the participants in any of these aspects of cognitive ability, nor did the app make any promise of improvement to cognitive performance through repeated play. Instead, the app seeks to measure the natural variation in performance on such tasks throughout the day [13] and in relationship to an individual's sleep-wake cycle [14]. The app also aims to utilise a large scale sample to answer the question of whether such variations are related to an individual's Morningness-Eveningness score – i.e. whether 'Larks' perform better earlier in the day than 'Owls' who perform better later on [15].

Here we present the in-game data from the app, report the broad performance of our cohort across the five tasks in relationship to the respective task literatures, any relationships between demographic factors and performance, and discuss broader issues of gamification and task design for use in app-based testing.

Methods

The OU Brainwave app was designed and created in collaboration with an external developer (Conjure Ltd, London, UK). Each of the games included in the app went through numerous rounds of development, with usability and participant engagement given equal weight to the essential factors of data validity and experimental design. The app was launched on Android and iOS mobile phone platforms in February of 2015. The launch was publicised through blog posts, and traditional media coverage. In addition, participants were encouraged to publicise the app through a built-in function to share a graph of their own results to social media.

Ethical approval for this study was obtained from the Open University human research ethics committee. Immediately upon downloading and opening the app, participants were presented with an informed consent statement which they were required to agree to via a tick box in order to continue using the app. Once consented, a unique participant ID for each participant was generated to link participant and future session data. Should the participant wish to withdraw their consent at a later date they could do so through a settings screen. Doing this deleted all participant data on the device and returned the participant to the opening screen of the app, where they had to agree again to the consent statement in order to use the app again. At no point was any personal or potentially identifying information collected from the participants.

Participants entered simple demographic information; sex (male/female) and age in years, although participants could choose not to answer either of these questions. Participants then completed the five item Morningness-Eveningness self-report questionnaire (MEQ) [16]. The MEQ is a well-established and validated research tool [13,17,18], and the 5 item variation of the original questionnaire was used here to move participants through onto the more interactive aspects of the app as quickly as possible. Using the original scoring of this implementation of the MEQ, participants are coded into one of five types ranging from "strongly morning type" to "strongly evening type". This result was shared on screen with the participant and, to encourage continued and repeated participation, they were then prompted to continue to the games in order to "see if your performance matches your belief".

The app also attempted to ameliorate the high attrition rates from which mobile phone apps suffer by displaying the participant performance graph only once the participant had completed three sessions. This was made clear to the participant each time they used the app until they had completed this requirement, at which point a graph of their performance, on each of the games and as an aggregate score, was shown. These graphs were designed in order to show the participant the variation in their performance on the tasks across the day and night, rather than to reveal their absolute performance levels. As such the performance values were normalised for each participant in order to highlight their best and worst scoring sessions. Accompanying the presentation of these graphs were icons encouraging the participant to share the image on social media.

At the beginning of each session – i.e. on each subsequent launching of the app – the participants were also asked up to three additional questions. A single item mood rating [19] was included at the start of every session, "How is your mood right now?" which they responded to via a Visual Analogue Scale (VAS) slider and if a session was the first on a given day, the participant was also asked what time they had woken up from sleep and how many hours of sleep they had had the previous night. Participants could opt to skip answering these questions and continue onto the games. Each session comprised of all five games, which were presented in a randomised order. Participants could choose to skip any game during a session, but were encouraged to complete them all through game-by-game results graphs of their own performance within the app. These graphs were only shown at all after 3 full sessions, to encourage a minimum level of engagement, and updated with each play after this point to promote continued play.

Games

"Hotspot" Action acquisition task

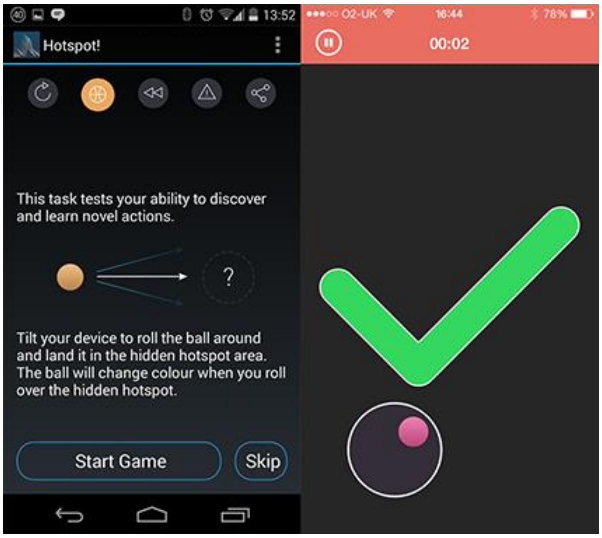

Figure 1 - Hotspot game instructions and play screen. Participants moved their device in order to roll an on-screen ball around and feedback was given then they successfully discovered the hidden 'hotspot'

The "Hotspot" game was a variation on an action discovery and acquisition task [20]. In this task participants must discover a target area by tilting their phone or tablet in order to roll an on-screen ball into a target area (figure 1). The target area is unmarked and no feedback is given until the target area is 'discovered' by the participant rolling the ball over the area, at which point the colour of the ball changes. The participant must then use this colour change to guide them in bringing the ball to rest within the target area. The difficulty of the task was adjusted in development by including a 100ms delay between success (i.e. entry into the target area) and feedback signal (i.e. colour change of the ball). The effect of delays of this type and magnitude is to increase the difficulty of the task [21], and was intended to prevent ceiling effects amongst the app participants. The game consisted of five attempts and each attempt presented a new, randomly chosen, target area covering 5% of the total space of the game arena, with the ball covering half of that area. To succeed in the game, participants had to keep the ball within the target area for 500ms of a 1 second window. Scores were allocated so that 50% of points available were awarded for finding the target and the remainder were apportioned according to milliseconds elapsed before the ball had remained within the target area for the required time.

"React" Persistent Vigilance task

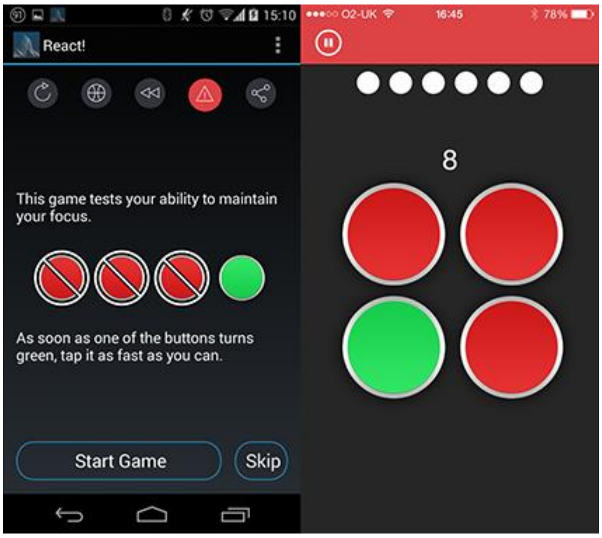

Figure 2 - React game instructions and play screen. Participants were required to hold their attention on the four red circles and press when one turned green. Auditory and visual feedback was given on correct and incorrect responses.

The "React" game was intended to be an implementation of the psychomotor vigilance task [22]. During the design process, it was decided to adjust the way the task was operationalised in the game to try to increase participant engagement. This was done by including a simple choice element, which was in addition to the reaction time task and not a standard part of the classic psychomotor vigilance task. Participants were presented with four large red, circular, buttons (figure 2). At a random interval between 2 and 7 seconds, one of the buttons changed colour to green and the participant had to tap on the appropriate button within a 600msec window. This was repeated eight times. Participant scores were essentially simple reaction time measures with scores reducing according to milliseconds elapsed before correct response was recorded, after a 100msec grace period. Responses made before the colour change, or incorrect button presses scored zero.

“Spin” Mental rotation task

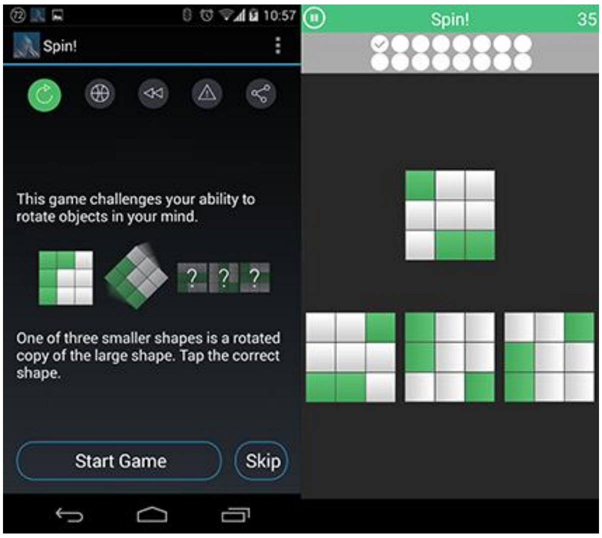

Figure 3 - Example instruction screen and in game screenshot. In the “Spin!” game, participants were required to match the test image with one of three options. Feedback was given in the form of ticks and crosses in the circles at the top of the screen, along with auditory feedback, and a timer was shown with the remaining time on the task

“Spin” was a gamified implementation of a spatial rotation task, using the stimulus set developed by Bethell-Fox & Shepard [23] as shown in figure 3. This stimulus set contained 18 possible patterns of filled squares within a 3x3 grid, avoiding excessive simplicity or difficulty, and rotational symmetry of pattern. Each pattern contained one, two or three groups of filled squares within the grid and while the original paper split these into levels of difficulty, all 18 were presented in a random order here within a given session to provide variation to the participant. Participants were presented with a large image of the target grid and had to correctly identify the rotated version of this grid from three alternatives presented below. The correct version was rotated at random by 90, 180 or 270 degrees and the incorrect options consisted of the test pattern reflected either vertically or horizontally. Participants had 45 seconds to make as many correct judgements as they could, up to a maximum of 18 and correct/incorrect auditory and visual feedback was given after each response.

"Super Snap" N-back analogue

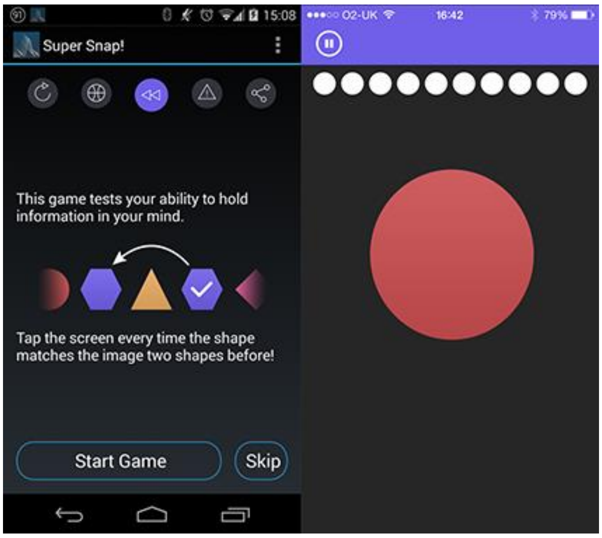

Figure 4 - Snap game instruction and play screen. Participants tapped on the shape on screen when it matched the shape shown two before. Auditory and visual feedback was given after each response, along with a tick or cross at the top of the gameplay window.

A simple implementation of the classic N-back task was used as a test of working memory (figure 4). The N-back task has a long history of use in studies of working memory [24] though see Kane [25] for a detailed discussion of the construct validity of the N-back. Here, a series of six brightly coloured shapes (circle, hexagon, rhombus, square, star, and triangle) were shown on screen and participants had to tap the screen to mark when the current shaped matched that shown two shapes ago. Each shape was presented on screen for 1.5 seconds against a blank, black, background with inter-item delays of 1.5 seconds. Participants were scored by the number of correct responses and the game continued until ten matches had been presented or ten responses (including false alarm incorrect responses) had been made. Participants started each session with a score of 60 and lost 6 points for each incorrect response or miss recorded.

"Track" Multiple object tracking

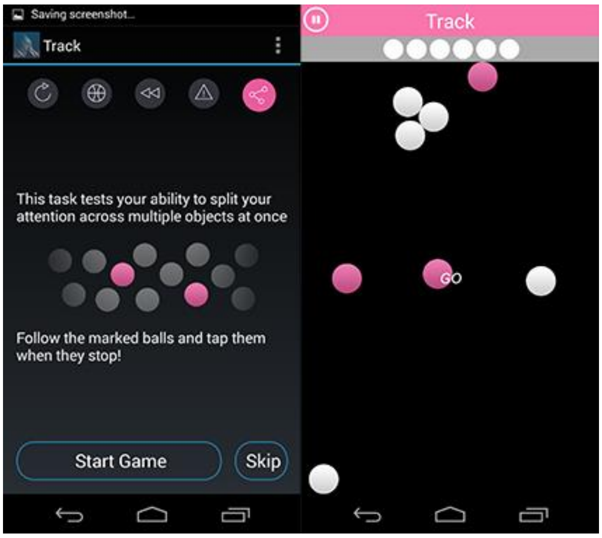

Figure 5 - Track game instruction and play screen. A subset of the balls onscreen were highlighted to the participant before the start of the trial, before reverting to white once the trial started and all balls began to move. After 5 seconds the balls stopped moving and participants were required to tap on the ones originally highlighted. Auditory and visual feedback was given after each trial along with a tick (for correct identification of all balls) or cross for each trial appearing along to top of the gameplay area.

The "track" game (figure 5) was a gamified version of a multiple-object tracking task [26]. In this task participants had to track the location of three members of an array of identical moving balls. Participants were first shown a static array of eight, nine or eleven balls, three of which were highlighted in pink rather than the white of the other balls. After a three second countdown the highlighted balls reverted to white and all the balls began moving on independent, randomly assigned, trajectories. The speed and direction of movement of each ball was adjusted randomly between each frame and collisions between balls or borders were handled such that no ball was ever overlapped or exceeded the playing area. The balls continued in motion for 5 seconds, after which time the entire array stopped and participants were instructed to tap the three balls which had been highlighted at the start of the trial. Two trials of each array size were shown, with the set sizes presented in increasing order. Participants were scored on number of balls correctly identified with non-responses counted as incorrect. Each correctly identified ball added a score of 2.5, so a maximum score of 45 across the 6 trials was possible.

Results

Demographics

Number of downloads and participants

The app launched in both the Apple and Android stores on January 15th 2015. Discovery of the app peaked in its launch month, with 4394 installations across January, with the expected drop off of installations broken only by smaller peaks in April 2015 and January 2016 – both probably due to further publicity activity. Again, like many apps OU brainwave found far more users on the Apple platform than Android – with roughly two thirds of the 15,890 users across the 18-and-a-half-month data collection window using Apple devices.

Separately, and far more importantly than the raw number of downloads is the number of active users of the app. As with many mobile phone apps, many downloaders either did not open the app or did not engage with the app sufficiently to be considered an active user. Of the 15,890 installations between application launch and July 31st 2016, 13,979 used the app sufficiently to upload some data to the server, meaning almost 2,000 downloaders did not go on to open the app after installation. 3661 users engaged with the app for a single session only – the most popular decision among downloaders - contributing at most demographic and MEQ data along with a single session's play to the dataset. Attrition among the remaining 10,319 was predictably steep with only 5756 users playing for more than three sessions, dropping to 1435 users at ten sessions or more. While, deliberately, no contact information was collected, so precluding any survey of the participants who stopped early, potential disincentives may have included technical issues, particularly with the Android app, the 'pestering' of the app notifications, or a perception that the demands of the app were too high. Just over one thousand users played 12 or more sessions and just over 100 played 30 or more sessions. Additionally, each user was free to play one, some or all five games during a given session so that while 3556 users completed five plays of any single game, only 2780 completed five plays of all five games. For this reason, later analyses are conducted at a game-by-game level and no overall performance measure is calculated.

Demographics of total participant cohort

Of the 13,979 total participants, 5517 self-reported as male (39.47%), 8231 as female (58.88%), with only 231 (1.65%) participants declining to answer the gender question. The self-report ages of all participants are shown in figure 6.

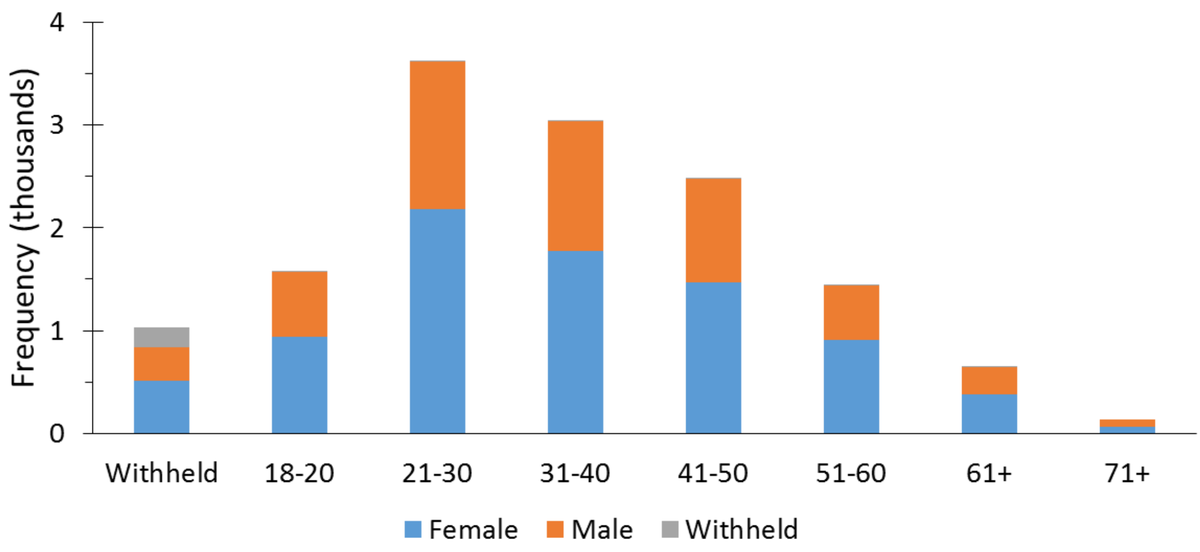

Figure 6 - Age of participants by gender

The distribution of reported age of participants shown in Figure 6 reveals a shortcoming of the implementation of self-report demographic data collection. While 1033 declined to answer the age question, an exceptionally large numbers of participants reported their age as 18 (965, compared with 405 for 27yrs, the next most common age). While this might well be an accurate figure, it could potentially be an artefact caused by the requirement for participants to confirm they are over 18 in order to use the app and play the games – a stipulation necessary for ethical approval. Eighteen was therefore the lowest available selectable age in the self-report question. Unless age controls on the downloading and installation of applications – controlled by the developer rather than end user (or parent) – become a viable option, future applications, especially particularly gamified ones, may wish to consider collecting and subsequently discarding (or filtering to not upload) data from particular age groups rather than attempting to exclude by self-report of age.

MEQ – Morningness Eveningness Questionnaire

The five-item Morningness-Eveningness self-report questionnaire [16] each participant completed produces a score of between 4 and 25, running between extreme evening type and extreme morning type. Respondents are traditionally then classified into five classes by their score (Definitely evening type – 4-7, moderately evening type – 8-11, neither type – 12-17, moderately morning type – 18-21, and definitely morning type – 22-25). From our original sample, 13752 participants filled in all sections of the MEQ survey. As in common in studies using the MEQ, around half of our overall sample scored within the “neither type”, central range of the MEQ (7172 participants, 52.2%). A further third of participants scored in one of the evening type categories (4584 participants, 33.3%) with the remaining 1996 (14.5%) scoring in the morning type categories.

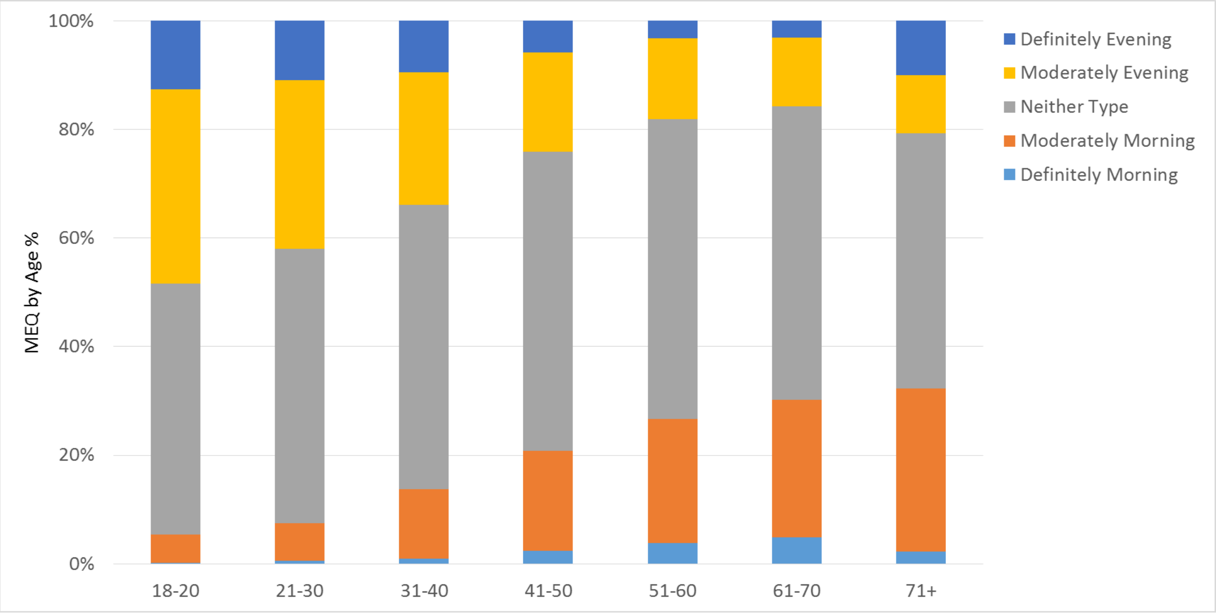

Figure 7 - MEQ scores by age

One of the stronger relationships usually found with the MEQ is that between age and Morningness [18,27] – greater age is associated with greater Morningness scores – and, as Figure 7 shows, we replicated that finding here. We found a statistically significant correlation between age and MEQ score for the participants who submitted both age and MEQ data ($r = 0.298$, $n = 12755$, $p < 0.001$). The greater absolute number of evening types than morning types in the dataset is almost certainly a result of this relationship being expressed in our cohort which has a skew toward younger participants.

As figure 8 shows, there was no significant difference between MEQ scores for males ($M = 13.20$, $SD = 4.00$) and females ($M=13.28$, $SD = 3.99$), ($t\ (13539) = -1.036$, $p = 0.30$). While many studies have reported a greater propensity for evening types in males compared to females [17], the lack of gender differences reported here is not an uncommon finding in the literature [18,28].

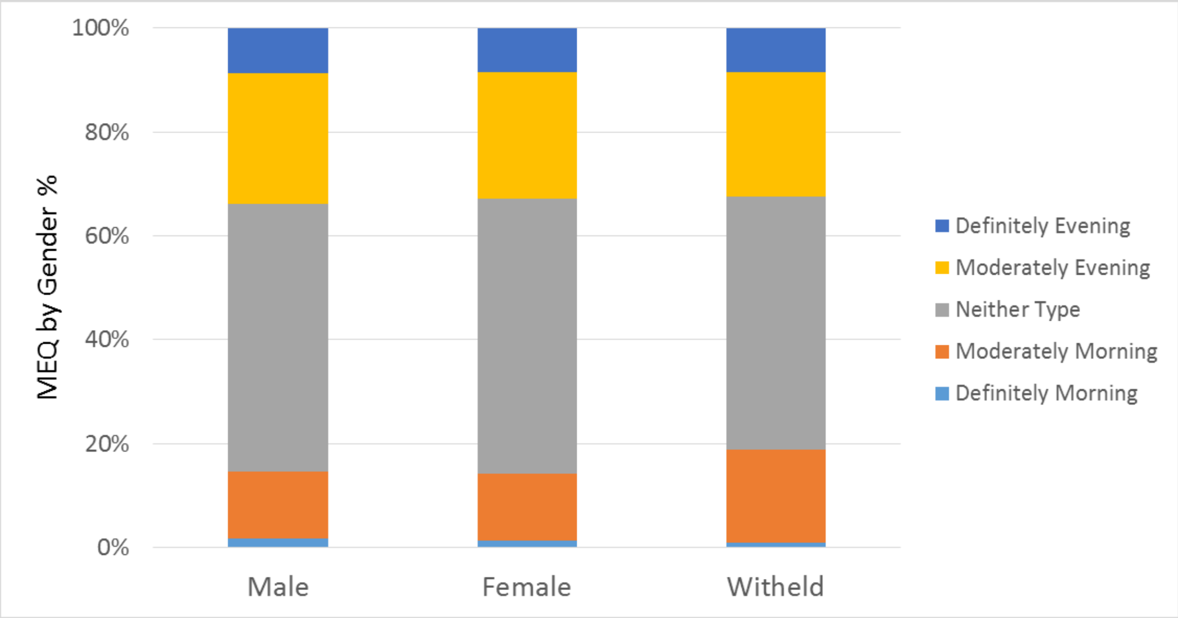

Figure 8 - MEQ score by gender.

Cognitive task results

Comparison of results to previous research and demographic effects

Because each participant could choose to play the separate games individually, any two participants will potentially have played any given game a different number of times. Additionally, the full participant set contains participants who did not play a particular game sufficiently to become familiar with it. This makes handling the data generated by the app very different from the usual data the tasks in OU Brainwave generate when conducted in a lab setting. To address the issue of participants who did not engage with a task sufficiently to even familiarise themselves with it, we implemented a cut off of a minimum of 4 plays of each individual game for any participant to be included in analysis for that game. The intention here was to remove participants who played no more than what would be considered a 'practice trial' set in a laboratory-based experiment. However, this still leaves variability in the number of measures per participant (in terms of sessions played) and the possibility that those participants who played more would register a higher mean score on each game. Therefore, the effect on score of demographic variables was analysed with number of plays as a co-variate. To remove outlier individuals, the most extreme 1% of average performance scores were identified and excluded before all analysis.

Gender

Gender effects were analysed using an ANCOVA with the mean score of the participant as the dependent variable, gender as a fixed factor, and number of plays as a covariate. To ensure gender and number of plays were not confounded, a t-test was conducted to confirm that there was no significant difference between the number of plays of a particular game by each gender. For each of the five games this test was non-significant. Table 1, below, shows the mean number of plays across gender for each game and the associated t-test statistic.

Game	N		Mean number of plays (Standard Deviation)		t-test statistic
	Male	Female	Male	Female	
Hotspot	1403	2640	8.07 (8.56)	8.36 (7.12)	t(4041) = -1.150, p = 0.250
React	1495	2840	8.09 (8.52)	8.32 (7.16)	t(4333) = -0.933, p = 0.351
Snap	1459	2773	8.13 (8.52)	8.42 (7.31)	t(4230) = -1.175, p = 0.240
Spin	1475	2784	8.16 (8.61)	8.40 (7.36)	t(4257) = -0.932, p = 0.351
Track	1457	2735	8.12 (8.57)	8.38 (7.30)	t(4190) = -1.038, p = 0.300

Table 1 - mean number of plays by game and gender

“Track” Multiple object tracking

A total of 4188 participants completed 4 or more sessions of the ‘Track’ game (1455 Male, 2733 Female). Previous studies in using multiple object tracking paradigms have shown an advantage for male participants (Valdes, Hines, Neill, & A., 2004) and we replicate that finding here. There was a significant, but small, effect of gender on performance after controlling for number of times beyond 4 that each participant played the game (F (1, 4185) = 11.047, p=0.001, η_ρ^2 = 0.003). Male participants scored on average an extra 1.2 points than female participants (Male M=38.9 SD = 12.0, Female M=37.7 SD = 11.7), though it should be noted that the variance in score accounted for by gender is very small (~0.3%) and only roughly half that accounted for by the significant effect of the number of plays beyond 4 by a particular participant (F (1, 4174) = 27.524, p<0.001, η_ρ^2 = 0.007).

“Super Snap” N-back analogue

4215 participants completed 4 or more sessions of the N-back analogue “Super Snap” (1449 Male, 2766 Female). We found no effect of gender on performance after controlling for number of plays beyond 4 (F (1, 4212) = 2.711, p=0.1, η_ρ^2 = 0.001), although number of plays beyond 4 was still found to significantly improve performance (F (1, 4227) = 127.06, p<0.001, η_ρ^2 = 0.03).

"React" Persistent Vigilance task

4241 participants completed 4 or more sessions of the "React" (1460 Male, 2781 Female). We found a small but significant effect of gender on performance after controlling for number of plays beyond 4 (F (1, 4238) = 15.993, p<0.001, η_ρ^2 = 0.004), accounting for 0.4% of the variance in mean score. Male participants scored an average 4 points more than female participants (Male M=409.0 SD = 29.6, Female M=405.1 SD = 29.9). Number of plays beyond 4 was not found to significantly improve performance (F (1, 4238) = 1.264, p=0.26).

"Spin" Mental rotation task

4219 participants completed 4 or more sessions of Spin (1455 Male, 2764 Female). We found a significant effect of gender on performance after controlling for the effect of number of plays beyond 4 (F (1, 4216) = 154.861, p<0.001, η_ρ^2 = 0.035), accounting for 3.5% of the variance in score. Male participants scored on average an extra 3 points than female participants (Male M=23.6 SD = 7.7, Female M=20.6 SD = 7.5). Number of plays beyond 4 was still found to have a significant, albeit smaller, benefit to performance (F (1, 4216) = 60.45, p<0.001, η_ρ^2 = 0.014).

"Hotspot" Action learning task

3991 participants completed 4 or more sessions of the N-back analogue "Super Snap" (1393 Male, 2598 Female). We found a small but significant effect of gender on performance after controlling for number of plays beyond 4 (F (1, 3988) = 90.59, p<0.001, η_ρ^2 = 0.022), with male participants scoring on average 3.7 points more than female participants (Male M = 31.33 SD = 12.25, Female M = 27.64 SD 12.03) and accounting for 2.2% of the variance in mean scores across the cohort. However, number of plays beyond 4 was also found to have a significant effect of similar size on performance (F (1, 3988) = 105.946, p<0.001, η_ρ^2 = 0.026).

In summary, of the five games all but the N-back task showed a significant effect of gender, controlling for the number of plays beyond 4. In each game which showed an effect, male participants scored higher (React, Track, Spin and Hotspot), with the strongest effect in the mental-rotation-based game (Spin).

Age effects

For every game, there was a significant positive correlation between the age of a participant and the number of plays (React N=4171, $r = 0.216$, Snap N=4081, $r = 0.221$, Spin N=4100, $r = 0.216$, Hotspot N= 3896, $r = 0.215$, Track N=4035, $r = 0.266$, all significant at $p<0.001$). Older participants played far more sessions than their younger counterparts, possibly due to the self-selecting nature of the cohort. Demographically, older people are less likely to either have a mobile phone or use a mobile phone for playing games [29], so it may be the case that for an older person, downloading the app represented a greater commitment to engage with the app than for a comparable younger downloader. With such a confounding relationship between number of plays and age, it was not appropriate to adopt the same approach - an ANCOVA using mean score for each participant - employed to analyse gender. Instead, a stratified approach was used to analyse the effect of age on performance controlling for number of plays. Rather than calculate mean scores from all of a participant's sessions, the mean score for each participant from only their 4th, 5th and 6th sessions was calculated. These early snapshots provide a measure of each participant's performance before they went on to complete their differing numbers of sessions and attain their eventual average performance level.

Using this measure, we then broke the participants into seven age groups by decade (<20, 20-29, 30-39, 40-49, 50-59, 60-69, & >70). This allowed us to analyse any potential age effects with an ANOVA, and compute the effect size for each game. We found a significant effect of age in all games except for Hotspot. The React, Snap, and Spin results (React (N=1596, $F(6, 1595) = 30.23$, $p<0.001$, $\eta^2 = 0.102$), Snap (N = 1627, $F(6, 1626) = 19.78$, $p<0.001$, $\eta^2 = 0.068$), Spin (N = 1640, $F(6, 1639) = 23.74$, $p<0.001$, $\eta^2 = 0.080$)) are all in line with previous findings which show strong negative associations between age and reaction time [30], age and working memory as measured by the N-back task [31], and age and mental rotation [32]. The track game, as an implementation of the multiple-object-tracking paradigm, might have been expected to similarly replicate a previously found age effect for multiple-object-tracking [33]. While we did find a significant main effect of age on multiple-object-tracking performance, this was a much smaller effect than those for react, snap and spin (N = 1616, $F(6, 1615) = 2.48$, $p= 0.022$, $\eta^2 = 0.009$). There was no relationship between age and performance on the action discovery task - the Hotspot game (N = 1519, $F(6, 1518) = 1.78$,

p=0.10, η^2 = 0.007). These analyses were also conducted with a correlational approach and the pattern of results was broadly similar.

Mood

Before starting each session of games, participants were presented with a single mood rating to which they responded to via a VAS slider. This additional data was collected to investigate the relationship between mood and cognition. Previous research has suggested a complex relationship between emotions, mood, and performance on cognitive tasks [34,35], with both beneficial and detrimental impacts on cognitive performance reported from a single, for example positive, mood induction [36]. Here, rather than induce a given mood, we simply record the participant's self-report of pre-existing mood, captured by a single item VAS measure (1 to 10, 10 being happiest) [19].

To be as inclusive as possible, all participants completing 4 or more plays of each game were included in this analysis, and collapsed across age and gender, giving group sizes of between 3988 (Hotspot) and 4331 (React) for each of the five games. No strong relationship between average mood and average performance of participant was found for any of the five games, with all Pearson correlation values being between 0 and 0.15. Similarly when the correlation coefficients for every participant were calculated individually for each game, no relationship was found, with the mean rho value being less than 0.02 in all cases, and for Snap, Spin, and Track not significantly different from 0. While future analysis may explore the possibility of non-monotonic relationships between mood and performance or potential differences in sub groups of the cohort, this first analysis suggests either that a single item VAS measure is insensitive to impactful changes of mood, or that mood and performance were unrelated on any task in the app.

Practice effects/learning curve analysis

The freedom to play each game an uncontrolled number of times by each participant and the likelihood that this would have an effect on an individual's performance, mean that either including number of plays as a covariate, or controlling this out in analysis is the most effective way to deal with the individual freedom of a gamified testing platform. However, it is still possible to visualise at the group level the effect of number of plays on performance. Plotting the mean scores for all participants in each game, except for the React game, shows typical practice effects as the participants familiarised themselves with the tasks. Practice effects are very common in psychological and psychometric testing and almost certainly reflect some combination of task familiarisation, development of the ability tested by the task, development of a strategy to complete the task as set and/or possibly reduced anxiety about the mechanics of the task [37].

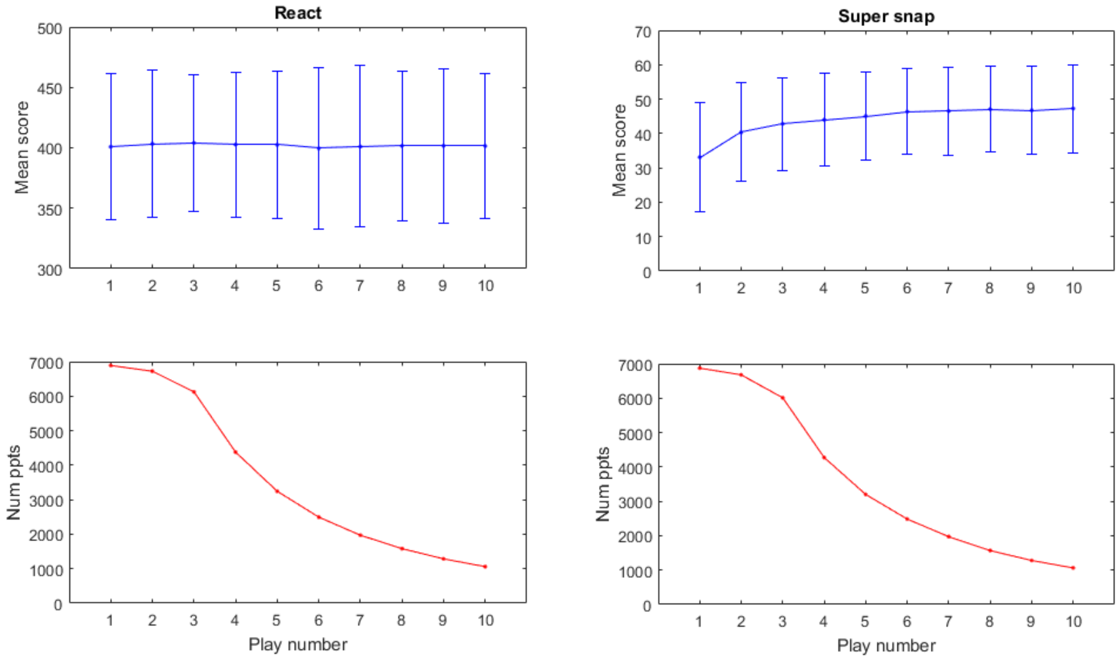

Figure 9 -

While four of the five games show typical practice effects - a stabilising of performance following an initial rise [38] - the React game, being essentially a very basic reaction time task, seems to have been too simple to produce any practice effect driven improvement in performance across participants' first few sessions (see figure 9), probably because participants immediately familiarised themselves with the task on first play and there are no effective strategies that can be adopted to improve performance. However, the variability of the cohort as a whole - in terms of inter-play interval, age, gender, MEQ etc. - mean that further group analysis of practice effects is unnecessary. The presence of the expected stabilising of performance after a number of plays reflects the intention of the app to measure variations in performance, rather than train or improve participants' abilities. Most importantly it means that future analysis of within individual factors - e.g. time of day of play - should have a stable performance level from which to contrast such changes.

Discussion

Success of app approaches

The large cohort collected by the OU Brainwave app, and moreover the repeated measurement of this cohort in quick, engaging gamified versions of classic and novel psychological tests, is another demonstration of the promise of mobile app-based testing [39]. While we deliberately did not collect more detailed demographic data, it is safe to say that with such a large sample, our testing cohort would have been extremely diverse compared to samples drawn from undergraduate participant pools that typify much laboratory-based research. This, along with the sheer size of the cohort tested, should mean that any reliable findings arising from this dataset are relatively robust and not hostage to cohort effects.

The usual caveats regarding the reliability of self-report data apply to our demographic and MEQ responses [40]. While this has resulted in the potential concern regarding the high number of participant's self-reporting their age as 18, the possibility remains that this is an accurate reflection of the participant cohort. Moreover, as a full data set, the cohort replicates a number of age-related findings, both in terms of the increased Morningness in older participants and reductions in cognitive task performance. The very low proportion of participants who withheld demographic information - only 231 participants, less than 2% of our total cohort, withheld either their age or gender information - is very encouraging for future mobile phone based research which can expect to have a high level of engagement from participants who download the app.

OU Brainwave is not the most downloaded research-focused app, and since its release a number of impressively large data sets have been collected and published using other app platforms [41] and online web-apps [42]. However, while OU Brainwave suffered from the same participant attrition as all apps, it recruited an impressively engaged cohort who repeatedly played the games (1400 playing more than 10 times) even though they did not vary or become more challenging with continued play. This suggests that some of the participant engagement features were successful.

A key intention in the development of OU Brainwave was to balance the demands of behavioural experiments, in terms of data validity and the operationalisation of the mechanism under study, against the enjoyment and engagement of the participant. The high levels of engagement of the participants who downloaded the app suggest that an in-game narrative, characters to interact with, or even an elaborate game environment may not be necessary. Studies directly manipulating the extent of gamification have reported a similar lack of effect of common gamification techniques on participant attrition [43]. The games, while offering dramatically shorter sessions than one would find in laboratory testing, did not deviate far from their experimental task heritage. Except for the React game, no significant change was made from the mechanics of the underlying psychological tasks, and the tasks were presented without cutesy preambles, fictional scenarios, or even in-game-rewards beyond the simple graphing of participant performance. The withholding of each participant's performance graphs until they had completed 3 sessions is likely to have had the effect of carrying more participants through the steepest part of the attrition and may have contributed to the longevity of the app for these participants. Similarly, the embedded ability to share one's own performance graph - which constantly updated with continued play - encouraged both the spread of the app and the continued engagement of the individual participant. Future experimental psychology apps should potentially focus on these features during app design as potentially highly effective, and simple, tools to encourage participation. On the other hand, the tendency of our older participants to contribute more data in terms of sessions played may suggest this group had a greater intrinsic motivation to engage with the app, or at least suggest that participants' engagement was a function of both intrinsic motivations and app features or in-game mechanics. Future studies may find it valuable to survey users during development in order to isolate the most valuable engagement features.

The inclusion of the full rMEQ [16], and the placing of it right at the start of the app experience meant that the Morningness data collected provides possibly the strongest finding of this early analysis. Our sample of over twelve and a half thousand adults revealed strong evidence for increased age correlating with Morningness – with older people being more likely to be moderate or strong morning types than younger people who, in turn are far more likely to report themselves as moderate or strong evening types. We do not replicate the finding for a similar tendency toward Morningness among female respondents compared to male, but as with the strong relationship found with age; this result is in line with previous studies.

Analysis of performance in the five cognitive tasks which make up the games of the OU Brainwave app showed that the app produced valid data and is sensitive enough to detect small but significant effects of both age and gender on cognition. We found small effects of gender in four of the five tasks (React, Track, Spin and Hotspot) and no effect in the other (Super Snap). In each case where a difference was found, the male participants scored slightly higher on average than the female participants. The largest difference was found in the Spin game, where gender accounted for 3.5% of variance in average score. The spin game is a direct implementation of a mental rotation task, a task which has previously been found to produce large, reliable gender effects [44]. While all the gender effects reported here are small, this could well be due to the freedom given to the participants and the resulting noise in the dataset. Further the unidirectional pattern of the gender effects reported here mean that an alternative explanation for these effect of platform - i.e. mobile phone app - rather than cognitive task, cannot be ruled out.

The impact of age on game score was much more pronounced than that of gender. Here we report a significant reduction in game scores for older compared to younger participants in all but the Hotspot game, and comparatively large effects in the Snap, Spin and React games where age accounted for 6.8%, 8% and 10.2%, respectively, of variation in average score. That we found our largest effect of age related decline in the task most heavily reliant on reaction time is of no surprise - increases in reaction time have long been associated with increasing age [45]. However, the evidence we found for the adverse effect of age on both mental rotation (Spin) and working memory (Super Snap), which involve more sophisticated constructs than simple reaction time, suggests that the app is indeed sensitive to fine grained differences in specific aspects of cognitive performance.

Future analysis will focus on the effect of time of day on performance. For example, those who report themselves as morning types can be compared against those who report themselves as evening types across each of the five tasks, though care will need to be taken to check and account for any bias induced by the self-reporting of MEQ before task performance began. This approach will enable the MEQ scores to be directly compared against task scores, on tasks whose sensitivity we now have some understanding of. Further, the recording of both hours spent sleeping in the previous night and time woken from sleep on each testing day from each participant will enable us to analyse the relative contribution of time spent awake, duration of preceding sleep, and time-of-day on any variation in cognitive performance.

The OU Brainwave app, with its cohort of ~14,000 active participants represents an exciting and rich dataset. The user-focussed features built into the app - extremely short testing durations, allowing participants to manage their own participation, engaging them through in-app feedback of their performance, and encouraging them to become an active part of the recruitment process by sharing their own performance - were largely very successful. The variability this approach introduced into the resulting performance data presented a challenge to data analysis. However, the replication of expected results, and the sensitivity of the app to group level differences in performance reported here all suggest that research apps which focus on user engagement and enjoyment, even at the expense of rigid and rigorous experimental protocols, produce valid and valuable data. Future data-collecting research apps may benefit from a similar focus on participants as users, not just data points.

Acknowledgements

The app development central to this project was supported by a grant from the Reed Foundation. Author contributions: MT designed and developed the app, analysed the Data and wrote the manuscript. JL analysed the data and contributed to the manuscript. DL & GP designed and developed the app and contributed to the manuscript.

References

1. Chetverikov, A., and Upravitelev, P. (2016). Online versus offline: The Web as a medium for response time data collection. Behav. Res. Methods *48*, 1086–1099. Available at: http://link.springer.com/10.3758/s13428-015-0632-x [Accessed December 21, 2016].
2. Barnhoorn, J.S., Haasnoot, E., Bocanegra, B.R., and van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *47*. Available at: http://link.springer.com/10.3758/s13428-014-0530-7 [Accessed December 21, 2016].
3. Brown, H.R., Zeidman, P., Smittenaar, P., Adams, R.A., McNab, F., Rutledge, R.B., and Dolan, R.J. (2014). Crowdsourcing for cognitive science--the utility of smartphones. PLoS One *9*, e100662. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0100662 [Accessed September 21, 2015].
4. Henrich, J., Heine, S.J., and Norenzayan, A. (2010). The weirdest people in the world? Behav. Brain Sci. *33*, 61–83. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20550733 [Accessed June 8, 2018].
5. Grady, C. (2012). The cognitive neuroscience of ageing. Nat. Rev. Neurosci. *13*, 491–505. Available at: http://www.nature.com/doifinder/10.1038/nrn3256 [Accessed March 27, 2017].
6. Bourne, L.E., Kole, J.A., and Healy, A.F. (2014). Expertise: Defined, described, explained. Front. Psychol. *5*, 186. Available at: http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00186/abstract [Accessed June

30, 2017].

7. Hoffman, R.R. (1996). How can expertise be defined?: Implications of research from cognitive psychology. In Exploring Expertise, R. Williams, W. Faulkner, and J. Fleck, eds. (Edinburgh: University of Edinburgh Press), pp. 81–100.

8. Williams, A.M., Ford, P.R., Eccles, D.W., and Ward, P. (2011). Perceptual-cognitive expertise in sport and its acquisition: Implications for applied cognitive psychology. Appl. Cogn. Psychol. *25*, 432–442. Available at: http://doi.wiley.com/10.1002/acp.1710 [Accessed June 8, 2018].

9. Kingstone, A., Smilek, D., and Eastwood, J.D. (2008). Cognitive Ethology: a new approach for studying human cognition. Br. J. Psychol. *99*, 317–40. Available at: http://www.ingentaconnect.com/content/bpsoc/bjp/2008/00000099/00000003/art00001;jsessionid=4an12landg8ef.alice [Accessed August 26, 2010].

10. Anderson, J.R. (2015). Cognitive psychology and its implications 8th Ed. (New York, NY, USA: WH Freeman/Times Books/Henry Holt & Co.).

11. Machery, E. (2010). Explaining why experimental behavior varies across cultures: A missing step in “The weirdest people in the world?” Behav. Brain Sci. *33*, 101–102. Available at: http://www.journals.cambridge.org/abstract_S0140525X10000178 [Accessed July 6, 2017].

12. Lumsden, J., Skinner, A., Woods, A.T., Lawrence, N.S., and Munafò, M. (2016). The effects of gamelike features and test location on cognitive test performance and participant enjoyment. PeerJ *4*, e2184. Available at: https://peerj.com/articles/2184 [Accessed October 31, 2016].

13. Schmidt, C., Collette, F., Cajochen, C., and Peigneux, P. (2007). A time to think: circadian rhythms in human cognition. Cogn. Neuropsychol. *24*, 755–89. Available at: http://dx.doi.org/10.1080/02643290701754158 [Accessed September 1, 2014].

14. Blatter, K., and Cajochen, C. (2007). Circadian rhythms in cognitive performance: methodological constraints, protocols, theoretical underpinnings. Physiol. Behav. *90*, 196–208. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17055007 [Accessed July 14, 2014].

15. Song, J., and Stough, C. (2000). The relationship between morningness–eveningness, time-of-day, speed of information processing, and intelligence. Pers. Individ. Dif. *29*, 1179–1190. Available at: http://www.sciencedirect.com/science/article/pii/S0191886900000027 [Accessed September 17, 2014].

16. Adan, A., and Almirall, H. (1991). Horne & Östberg morningness-eveningness questionnaire: A reduced scale. Pers. Individ. Dif. *12*, 241–253.
17. Tonetti, L., Fabbri, M., Martoni, M., and Natale, V. (2011). Season of Birth and Sleep-Timing Preferences in Adolescents. Chronobiol. Int. *28*, 536–540. Available at: http://www.tandfonline.com/doi/full/10.3109/07420528.2011.590261 [Accessed June 12, 2017].
18. Paine, S.-J., Gander, P.H., and Travier, N. (2006). The Epidemiology of Morningness/Eveningness: Influence of Age, Gender, Ethnicity, and Socioeconomic Factors in Adults (30-49 Years). J. Biol. Rhythms *21*, 68–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16461986 [Accessed June 12, 2017].
19. Folstein, Marshal F., Luria, R. (1973). Reliability, validity, and clinical application of the visual analogue mood scale. Psychol. Med. *3*, 479–486. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4762224 [Accessed December 22, 2016].
20. Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., Gurney, K., and Redgrave, P. (2012). A Novel Task for the Investigation of Action Acquisition. PLoS One *7*, e37749. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037749 [Accessed July 27, 2015].
21. Walton, T., Thirkettle, M., Redgrave, P., Gurney, K.N., and Stafford, T. (2013). The discovery of novel actions is affected by very brief reinforcement delays and reinforcement modality. J. Mot. Behav. *45*, 351–60. Available at: http://www.tandfonline.com/doi/abs/10.1080/00222895.2013.806108 [Accessed October 29, 2014].
22. Basner, M., Mollicone, D., and Dinges, D.F. (2011). Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. Acta Astronaut. *69*, 949–959.
23. Bethell-Fox, C.E., and Shepard, R.N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. J. Exp. Psychol. Hum. Percept. Perform. *14*, 12–23. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.14.1.12.
24. Oberauer, K. (2005). Binding and Inhibition in Working Memory: Individual and Age Differences in Short-Term Recognition. J. Exp. Psychol. Gen. *134*, 368–387. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.134.3.368 [Accessed June 10, 2016].
25. Kane, M.J., Conway, A.R.A., Miura, T.K., and Colflesh, G.J.H. (2007). Working

Memory , Attention Control , and the N -Back Task: A Question of Construct Validity. J. Exp. Psychol. Learn. Mem. Cogn. *33*, 615–622. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.33.3.615 [Accessed June 9, 2016].

26. Pylyshyn, Z.W., and Storm, R.W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spat. Vis. *3*, 179–97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3153671 [Accessed August 24, 2014].

27. Adan, A., Archer, S.N., Hidalgo, M.P., Di Milia, L., Natale, V., and Randler, C. (2012). Circadian typology: a comprehensive review. Chronobiol. Int. *29*, 1153–75. Available at: http://informahealthcare.com/doi/abs/10.3109/07420528.2012.719971 [Accessed August 7, 2014].

28. Zimmermann, L.K. (2011). Chronotype and the Transition to College Life. Chronobiol. Int. *28*, 904–910. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22080735 [Accessed June 12, 2017].

29. Ofcom (2015). Adults' media use and attitudes Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0014/82112/2015_adults_media_use_and_attitudes_report.pdf [Accessed July 14, 2017].

30. Verhaeghen, P., and Salthouse, T.A. (1997). Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. Psychol. Bull. *122*, 231–249. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9354147 [Accessed April 7, 2017].

31. Kirchner, W.K. (1958). Age differences in short-term retention of rapidly changing information. J. Exp. Psychol. *55*, 352–358. Available at: http://www.ncbi.nlm.nih.gov/pubmed/13539317 [Accessed April 25, 2017].

32. Berg, C., Hertzog, C., and Hunt, E. (1982). Age differences in the speed of mental rotation. Dev. Psychol. *18*, 95–107. Available at: http://content.apa.org/journals/dev/18/1/95 [Accessed April 7, 2017].

33. Trick, L.M., Perl, T., and Sethi, N. (2005). Age-Related Differences in Multiple-Object Tracking. Journals Gerontol. Ser. B Psychol. Sci. Soc. Sci. *60*, P102–P105. Available at: http://psychsocgerontology.oxfordjournals.org/content/60/2/P102.short [Accessed October 5, 2016].

34. Oaksford, M., Morris, F., Grainger, B., and Williams, J.M.G. (1996). Mood , Reasoning , and Central Executive Processes. J. Exp. Psychol. *22*, 476–492. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.22.2.476 [Accessed April 25,

2017].
35. Clore, G.L., and Huntsinger, J.R. (2007). How emotions inform judgment and regulate thought. Trends Cogn. Sci. *11*, 393–399. Available at: http://www.sciencedirect.com/science/article/pii/S1364661307001908 [Accessed April 25, 2017].
36. Gray, J.R. (2001). Emotional modulation of cognitive control: approach-withdrawal states double-dissociate spatial from verbal two-back task performance. J. Exp. Psychol. Gen. *130*, 436–452. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.130.3.436 [Accessed April 25, 2017].
37. Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., and Moriarty Gerrard, M.O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. J. Appl. Psychol. *92*, 373–385. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.92.2.373.
38. Falleti, M.G., Maruff, P., Collie, A., and Darby, D.G. (2006). Practice Effects Associated with the Repeated Assessment of Cognitive Function Using the CogState Battery at 10-minute, One Week and One Month Test-retest Intervals. J. Clin. Exp. Neuropsychol. *28*, 1095–1112. Available at: http://www.tandfonline.com/doi/abs/10.1080/13803390500205718 [Accessed June 12, 2017].
39. Lumsden, J., Edwards, E.A., Lawrence, N.S., Coyle, D., and Munafò, M.R. (2016). Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. JMIR serious games *4*, e11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27421244 [Accessed September 20, 2016].
40. Chan, D. (2008). So why ask me? Are self-report data really that bad? In Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences (Routledge), pp. 309–336. Available at: http://www.tandfebooks.com/isbn/9780203867266.
41. Hunt, L.T., Rutledge, R.B., Malalasekera, W.M.N., Kennerley, S.W., and Dolan, R.J. (2016). Approach-Induced Biases in Human Information Sampling. PLoS Biol. *14*, e2000638. Available at: http://dx.plos.org/10.1371/journal.pbio.2000638 [Accessed June 13, 2017].
42. Stafford, T., and Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. Psychol. Sci. *25*, 511–8. Available at:

http://pss.sagepub.com/content/25/2/511.abstract [Accessed September 5, 2015].

43. Lumsden, J., Skinner, A., Coyle, D., Lawrence, N., and Munafo, M. (2017). Attrition from Web-Based Cognitive Testing: A Repeated Measures Comparison of Gamification Techniques. J. Med. Internet Res. *19*, e395. Available at: http://www.jmir.org/2017/11/e395/ [Accessed March 28, 2018].

44. Masters, M.S., and Sanders, B. (1993). Is the gender difference in mental rotation disappearing? Behav. Genet. *23*, 337–341. Available at: http://link.springer.com/10.1007/BF01067434 [Accessed September 18, 2014].

45. Der, G., and Deary, I.J. (2006). Age and sex differences in reaction time in adulthood: Results from the United Kingdom Health and Lifestyle Survey. Psychol. Aging.