Behavior Research Methods
https://doi.org/10.3758/s13428-021-01597-3

Cognitive test scores vary with choice of personal digital device

Eliza Passell[1] · Roger W. Strong[1,2] · Lauren A. Rutter[3] · Heesu Kim[1,4] · Luke Scheuer[1] · Paolo Martini[5] · Liz Grinspoon[1] · Laura Germine[1,2,5]

Accepted: 12 April 2021

Abstract

Mobile- and web-based psychological research are a valuable addition to the set of tools available for scientific study, reducing logistical barriers for research participation and allowing the recruitment of larger and more diverse participant groups. However, this comes at the cost of reduced control over the technology used by participants, which can introduce new sources of variability into study results. In this study, we examined differences in measured performance on timed and untimed cognitive tests between users of common digital devices in 59,587 (Study 1) and 3818 (Study 2) visitors to TestMyBrain.org, a web-based cognitive testing platform. Controlling for age, gender, educational background, and cognitive performance on an untimed vocabulary test, users of mobile devices, particularly Android smartphones, showed significantly slower performance on tests of reaction time than users of laptop and desktop computers, suggesting that differences in device latency affect measured reaction times. Users of devices that differ in user interface (e.g. screen size, mouse vs. touchscreen) also show significant differences ($p < 0.001$) in measured performance on tests requiring fast reactions or fine motor movements. By quantifying the contribution of device differences to measured cognitive performance in an online setting, we hope to improve the accuracy of mobile- and web-based cognitive assessments, allowing these methods to be used more effectively.

Keywords Web-based · Cognitive · Device · Digital · Reaction time

Introduction

Mobile- and web-based studies are a valuable addition to the set of tools available for psychological research. Digital cognitive assessments allow researchers to recruit participants on a broader scale than would be possible in an individual lab, by reducing logistical barriers such as travel to and from a research lab, enabling larger sample sizes with no need for additional recruitment efforts (Germine et al., 2012; Kraut et al., 2004). Research using personal digital devices for administering study measures has also proven to produce more diverse participant groups than are used in most conventional psychological research (Gosling & Mason, 2015). Computer-based measures can also provide types of data that traditional metrics do not, such as the time it takes participants to respond to individual survey questions (Buchanan, 2007). By allowing researchers to draw on a larger, more diverse, and more representative sample of the general population and obtain more precise data, online research using digital devices could lead to increased scientific rigor in the fields of psychology and cognitive science. In addition, making studies broadly available online and through participants' own devices makes it easier for participants to learn about and engage with research in psychology, opening up channels of communication between researchers and the general public. Research participation can have significant educational benefits for those learning about psychology; citizen science through online research could make these benefits available to all (de Liaño et al., 2012).

However, these benefits of research using personal digital devices come at the cost of control over the setting in which participants complete the study session (Nosek et al., 2002; Reips, 2000). This loss of control is particularly significant in relation to the specific technology used by study participants

✉ Eliza Passell
epassell@mclean.harvard.edu

[1] Institute for Technology in Psychiatry, McLean Hospital, Oaks Building, 115 Mill Street, Belmont, MA 02478, USA

[2] Department of Psychiatry, Harvard Medical School, Boston, MA, USA

[3] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

[4] Department of Neuroscience, Wellesley College, Wellesley, MA, USA

[5] The Many Brains Project, Belmont, MA, USA

Published online: 05 May 2021

(Woods et al., 2015). While a lab-based study allows researchers to provide any necessary digital devices and internet connection (and thus ensure that they remain consistent between all participants), participants in online research can use a wide variety of devices, including desktop and laptop computers, smartphones, tablets, and e-readers (De Bruijne & Wijnant, 2013). These devices can differ widely in hardware, software, and user interface, introducing uncontrolled and undetected variation into the experimental procedure (Germine et al., 2019; Woods et al., 2015), which may impact results.

One potential source of device variation that could affect data in psychological research, and particularly research involving reaction time, is device latency (Pavlovych & Gutwin, 2012). The latency of a device refers to the time necessary for a user's input (such as a mouse click or a touch on a touchscreen) to be registered by the computer. While device latencies for most commonly used technologies are short enough to be unnoticeable for most users, they can have a significant impact on time-dependent cognitive tests, such as tests of reaction time. The average simple reaction time has been shown to be in the range of 200–300 ms, with time increasing with age (Jain et al., 2015; Wilkinson & Allison, 1989; Woods et al., 2015). Therefore, variation in device latency, which usually ranges from 50 to 200 ms for touchscreen devices and 15 to 35 ms for modern desktop computers, has the potential to distort measured reaction times, causing true inter-individual differences in performance to be drowned out by increasing variation irrelevant to neurocognitive ability (Henze et al., 2016; Pavlovych & Gutwin, 2012). For example, some studies of the difference in reaction time between schizophrenic and non-schizophrenic subjects show differences of approximately 100 ms—small enough to be affected by the differences in measurement caused by device latency (Ngan & Liddle, 2000). Thus, without a way to control for variations in device latency, one cannot adapt reaction time research to an online format without running the risk of conflating device variation with actual variation in participant performance.

Measures of reaction time are a key instrument for research in psychological science. Despite its simplicity, reaction time provides a robust measure of cognitive function; performance on tests of simple reaction time (the time between the presentation of a cue stimulus and the completion of a response) and related tasks are correlated with a broad range of general intelligence measures, including less time-dependent measures based on arithmetic and vocabulary knowledge (Luciano et al., 2003). In addition, cognitive processing speed as measured by reaction time may explain the relationship between general intelligence and life outcomes; a study of 898 adults followed from ages 56 to 70 showed that the established relationship between general intelligence and all-causes mortality is no longer significant after controlling for reaction time (Deary & Der, 2005). As a measure of brain function, reaction time is associated with cognitive impairment across disorders, including schizophrenia, Parkinson's disease, and traumatic brain injury (Evarts et al., 1981; Hetherington et al., 1996; Ngan & Liddle, 2000). Reaction time can also be used as a tool to study attitudes, as in the Implicit Association Test, which measures differential association of two target concepts with an attribute by comparing the speed with which participants can pair each concept with the attribute of interest (Greenwald et al., 1998; Nosek et al., 2005). The information that can be obtained from reaction time is not limited to measurements of speed; older adults and people with cognitive impairments frequently show reaction times that are not only slower than the population average, but also more variable from trial to trial (Gorus et al., 2008; Hultsch et al., 2002). Typical reaction times on a test of simple reaction time are approximately 200–300 ms, so the 50–200 ms of delay contributed by device latency could be greater than the true variation between individuals (Jain et al., 2015; Ngan & Liddle, 2000; Wilkinson & Allison, 1989). The versatility of reaction time as a tool to provide insight into cognitive processes makes its accurate measurement vital for a variety of fields of research.

Because tests of reaction time and other timed tasks that take place on the scale of milliseconds are such vital tools in psychological research, it is crucial that, in taking advantage of the many benefits of digital cognitive assessment, we do not lose the ability to measure cognition accurately. In this study we use a large, diverse online sample to characterize the extent to which variations between digital devices used by participants affect their scores on reaction time-based tests. Based on our understanding of device latency, we predict that users of different digital devices will show differences in observed performance on cognitive tests based on reaction time. We expect that some of these differences will be associated with cognitive and demographic factors that vary with device ownership—for example, tablet users tend to be older than users of other devices. However, we expect that device-related differences will persist after controlling for demographic and cognitive differences measured by device-robust tasks (e.g. untimed tasks with simple stimuli). Because true reaction times scale with the cognitive complexity of the task, while device latencies do not, we expect that device-related differences will be the largest (as a proportion of overall variability) in tasks with short reaction times. In contrast, we expect smaller relative contributions of device variability on more complex tests that produce longer reaction times. Finally, we suggest methods to reduce potential device-related confounds.

Study 1: Differences in reaction time by device

In Study 1, we analyzed existing data on cognitive test performance and device characteristics from a large sample of visitors to TestMyBrain.org. We looked at each of the five most frequently used classes of devices on timed and untimed cognitive tests.

Methods

Participants

Study 1 was based on large samples of data collected across a range of tasks from 2014 to 2019 through the TestMyBrain.org platform. Our participants consisted of 59,587 volunteers who visited TestMyBrain.org, a citizen science website that allows people to participate in cognitive research studies in exchange for individualized feedback on their performance after each study. Data from TestMyBrain.org have been shown to be of similarly high quality as data gathered in a lab or clinic setting (Chaytor et al., 2020; Germine et al., 2012) and replicate findings from traditional lab or clinical settings (Hartshorne & Germine, 2015; Halberda et al., 2012; Germine et al., 2011). Before starting a task, participants read a study information sheet and gave informed consent. After completing the task, participants received feedback on their results relative to other individuals who had completed the same test. Consent and study procedures were reviewed by the Harvard Committee on the Use of Human Subjects.

Participants ranged in age from 18 to 80 years, with a mean age of 35.92 years (*SD* = 15.23). The sample contained similar numbers of male and female participants (48.56% male, 51.24% female, 0.20% non-binary/genderqueer). While the TestMyBrain platform is open to participants in all parts of the world, we limited our sample to participants who reported speaking English as a native language since one of our measures assessed English vocabulary knowledge. Further information on the demographic characteristics of the sample can be found in Table 1.

This study included data from all participants who completed one or more of the simple reaction time, digit symbol matching, or vocabulary (described below) tasks on TestMyBrain. These tasks were all available for participants to take through the online platforms for multiple periods of time between 2014 and 2019. These three tests were chosen in order to examine the effects of device latency and other device variability on both short and long reaction times. The simple reaction time task is intended to measure a participant's fastest possible response, while the digit symbol matching task is a speeded task that produces longer reaction times. The vocabulary test is not scored based on reaction time, and was included in order to compare the effects of device differences on timed and untimed tests. Participants whose devices could not be identified from user agent strings were excluded from analysis. Of 59,587 participants who completed any of these tests on an identifiable device, 30,747 completed simple reaction time, 17,507 completed digit symbol matching, and 21,519 completed vocabulary.

Measures

Simple reaction time This simple reaction time test requires participants to respond to a green square cue after a waiting period indicated by a red square cue. They are asked to respond by pressing the space bar or touching their touchscreen as quickly as they can. The time between the two cues varies between 700 and 1500 ms, with delays sampled based on an exponential random distribution from this range. There is also a consistent 700 ms delay between a participant's response in one trial and the presentation of the wait cue in the next trial. Participants complete 30 trials. This task is a measure of basic psychomotor response speed and has been used in previous studies (McLean et al., 2019; Passell et al., 2019). Notably, this task produces relatively short reaction times; the median reaction time in our sample was 303.86 ms (*SD* = 68.94 ms).

Digit symbol matching The digit symbol matching task requires participants to use a table of digit and symbol pairings present on the bottom of the screen throughout the duration of the task to indicate which of three numbers (1, 2, or 3) corresponds to a symbol shown on screen. Symbols are simple black and white icons. Each number is paired with three symbols in the key, for nine pairings in total. The pairings of symbols and digits remain consistent throughout the task. Participants who use laptop or desktop computers with keyboards respond by pressing the 1, 2, or 3 key on the keyboard, while participants using touchscreen-based devices respond by touching buttons on the screen labeled with the 3 digits. After a participant responds to a trial, there is a 250 ms delay before the next symbol is presented. This task was adapted from the classic Wechsler Adult Intelligence Scale (WAIS) digit symbol coding task (Drozdick et al., 2012) and adapted for web/mobile administration. This task is a measure of processing speed and short-term memory (Joy et al., 2004) and has been used in previous studies as a measure of processing speed across the lifespan (Hartshorne & Germine, 2015; McLean et al., 2019). This task produces longer reaction times than the simple reaction time task; the median reaction time on this test in our sample was 1360.03 ms (SD = 354.68). To best mimic the WAIS task, participants are given a time limit of 90 seconds to correctly answer as many trials as they can.

Vocabulary We used a 20-item test of English vocabulary, with words increasing in difficulty over the course of the test. In this test, the participant sees a word and has to select which of five words is closest in meaning to the target word. This test was included as a measure of cognitive ability that is not based on differences in reaction time, and thus would be expected to be relatively unaffected by device differences. This test has

Table 1 Study 1 participant characteristics and analysis sample sizes

Variable [range]	Analysis N	*n* (%)	*M* (*SD*)
Age [18–80]	59,587		35.92 (15.23)
Gender	59,119		
Male		28,709 (48.56)	
Female		30,291 (51.24)	
Non-binary/Genderqueer		179 (0.20)	
Education*	46,701		
Primary school		10 (0.02)	
Middle school		482 (1.03)	
High school		6024 (12.90)	
Some college		13,291 (28.46)	
Technical training/associate's degree		347 (0.74)	
Bachelor's degree		13,969 (29.91)	
Graduate school		12,578 (26.93)	

*Self-reported highest level of education completed

been used in previous studies as a measure of crystallized cognitive ability across the lifespan (Hartshorne & Germine, 2015; McLean et al., 2019; Passell et al., 2019). Vocabulary has been shown to correlate significantly, however, with performance on tests of processing speed and with overall cognitive performance (Luciano et al., 2004; Crawford, 1989).

Data analysis

All data analysis was performed in R (R Core Team, 2012). In addition to base R functions, we used the effsize R package to compute effect sizes and the uaparserjs R package to parse user agent information in order to identify digital devices (Rudis et al., 2016; Torchiano, 2019). We examined differences in reaction times by device group using linear mixed-effects models. In these models, the reaction time on each trial was treated as a continuous dependent variable and was scaled to have a mean of 0 and a standard deviation of 1, with results reported in standard deviation units. Participant ID and trial number were treated as random effects (random intercepts), allowing us to account for participants of different performance levels and trials of different difficulty levels. Device group, age (including quadratic and cubic effect of age), gender, age-gender interaction, and educational background were treated as fixed effects. Differences in vocabulary performance by device group were examined using a logistic mixed-effects model; accuracy on each trial was treated as a categorical (correct/incorrect) dependent variable, with participant ID and trial number treated as random effects and device group, age (including quadratic and cubic effect of age), gender, age-gender interaction, and educational background treated as fixed effects. In addition to these mixed-effects models, we also fit linear regression models for participant-level performance (median reaction time and vocabulary accuracy) with age, gender, and educational background as covariates; effect size was computed based on residuals from these models using Cohen's *d*.

To maintain data quality, we excluded participants whose performance indicated that they were not completing the task as intended. In reaction time tasks, we excluded individual trials with impossibly short response times and instead imputed the participant-level mean reaction time for those trials. The threshold for trial-level exclusions was set at 110 ms for simple reaction time and 600 ms for digit symbol matching. For simple reaction time, this threshold was determined based on established physiological limitations. For digit symbol matching, the threshold was based on examination of reaction times where accuracy consistently falls to chance level. We also excluded all trials in which the participant did not make a response within the time limit (2 seconds for simple reaction time; 10 seconds for digit symbol matching). If a participant had more than 10% of their reaction times outside of this range on simple reaction time (110–2000 ms) or digit symbol matching (600–10,000 ms), their data on the test in question were excluded from analysis. A total of 1175 participants (3.82%) had their simple reaction time data removed, while 350 participants (2.00%) had their digit symbol matching data removed. These exclusions were mostly due to fast responding (94.8% of outlying trials for excluded simple reaction time participants were below the minimum RT cutoff, while 99.5% of outlying trials for excluded digit symbol matching participants were below the minimum RT cutoff). In addition, participants who chose the same response (e.g. repeatedly selecting 3) 90% of the time or more in digit symbol matching had their data from this test excluded; of the participants who were not excluded for outlying reaction times, only two additional participants (0.0001%) had their data excluded based on this criterion. Vocabulary data were excluded if data showed that a participant had completed more or less than the expected 20 trials, since this was an indicator of a technical problem or participant error (e.g., reloading the page or returning to previous pages during the test). A total of 198 participants had vocabulary data removed based on this criterion.

Results

Effect of demographic differences in device use

The participant groups that completed tests on each device differed from one another in demographic characteristics as well as cognitive test performance (see Table 2). A one-way analysis of variance (ANOVA) showed a significant relationship between device choice and age ($F(4, 59{,}582) = 499.3$, $p < 0.001$, $\eta^2 = 0.03$); users of Macintosh, iPhone, and Android devices were younger, on average, than users of Windows devices, while users of iPad tablets had a higher average age than all other groups (see Table 2 for mean age of each group). A Games-Howell post hoc test showed that differences between all groups were significant. Device groups also differed by gender, with users of mobile devices and Macintosh laptop/desktop devices more likely to be female than users of Windows devices (see Table 2 for gender distribution of each device group). A chi-square test of independence showed that these differences were significant ($X^2(8) = 995.43$, $p < 0.001$, standardized mean difference effect size = 0.26); standardized residuals can be seen in Table 3. A one-way ANOVA also showed significant differences in educational background between different device groups ($F(4, 46{,}696) = 268$, $p < 0.001$, $\eta^2 = 0.02$); users of Macintosh devices had completed the most years of education on average and users of Android devices had completed the fewest (see Table 3 for educational background of each device group). Running a Games-Howell post hoc test, we found that differences between all groups were significant.

Age, gender, and educational background are known to predict performance on both timed and untimed cognitive tests. Previous research has shown that reaction time typically becomes slower over the course of the adult lifespan (Dykiert et al., 2012; Der & Deary, 2006). This is consistent with what we found in our data, which showed a significant positive correlation between age and latency in simple reaction time ($r = 0.25$, 95% CI [0.24, 0.26], $p < 0.001$). There is also evidence of gender differences in reaction time, with male participants showing faster reaction time on average (Dykiert et al., 2012); this pattern was also demonstrated in our simple reaction data ($t = -27.06$, $p < 0.001$). Thus, differences in observed performance between users of different devices could reflect differences in the demographic characteristics of people who choose each device, as well as the effects of the devices themselves. For this reason, we looked at the association between device type and cognitive test performance before and after controlling for age and gender.

Differences in performance by device

Mixed-effects linear regression models (Windows used as comparison group) showed that device type was associated with reaction time on both simple reaction time and digit symbol matching after controlling for age, gender, and educational attainment. This effect is greatest on simple reaction time, with Android, iPhone, and iPad devices showing longer reaction times than Windows devices (Android: $\beta = 0.56$, SE = 0.01, $t = 53.07$, $p < 0.001$; iPhone: $\beta = 0.21$, SE = 0.01, $t = 17.34$, $p < 0.001$; iPad: $\beta = 0.23$, SE = 0.01, $t = 16.02$, $p < 0.001$) and Macintosh laptop/desktop devices showing slightly shorter reaction times than Windows devices ($\beta = -0.07$, SE = 0.01, $t = -7.20$, $p < 0.001$). In the digit symbol matching test data, reaction times from users of Android and iPhone devices were significantly longer than those from users of Windows devices (Android: $\beta = 0.13$, SE = 0.01, $t = 9.45$, $p < 0.001$; iPhone: 0.09, SE = 0.01, $t = 5.81$, $p < 0.001$), while Macintosh users showed slightly shorter reaction times ($\beta = -0.07$, SE = 0.01, $t = -6.45$, $p < 0.001$). iPad users did not show significant differences in reaction time from Windows users ($\beta = -0.02$, SE = 0.02, $t = -1.18$, $p = 0.24$). On the untimed vocabulary test, mixed-effects logistic regression (also controlling for age, gender, and educational attainment) showed that use of Macintosh laptop and desktop devices

Table 2 Demographic differences in device use*

Device**	*N*	Mean age (SD)	Percent female	Mean years of education (SD)
Android	6593	33.23 (13.17)	55.01	14.72 (2.18)
iPhone	4678	31.99 (11.96)	61.39	15.01 (2.15)
iPad	3456	45.60 (15.85)	62.41	15.52 (2.21)
Macintosh	10,779	35.40 (15.40)	56.61	15.86 (1.96)
Windows	34,081	36.17 (15.45)	45.58	15.32 (2.14)

*Demographic statistics were computed based only on participants who successfully completed all tests and whose data were not removed for quality control

**Android devices include both mobile phones and tablets; Windows devices include tablet, laptop, and desktop computers

Table 3 Standardized residuals from chi-square test of independence of gender and device

	Windows	Android	iPad	iPhone	Macintosh
Female	−29.75	7.14	14.00	15.08	13.37
Non-binary/genderqueer	−3.35	2.29	1.60	2.61	−0.36
Male	30.05	−7.35	−14.15	−15.32	−13.34

was associated with correct responses compared to Windows devices (β = 0.30, SE = 0.03, z = 11.61, p < 0.001), while use of Android devices was associated with incorrect responses (β = −0.33, SE = 0.05, z = −7.06, p < 0.001); other devices did not show a significant association with accuracy (iPhone: β = −0.10, SE = 0.06, z = −1.84, p = 0.07; iPad: β = −0.01, SE = 0.05, z = −0.26, p = 0.87). The fact that differences persist for both timed test performance and vocabulary performance after controlling for age, gender, and education suggests that some of these differences may be due to some additional differences (other than age, gender, and education) in the makeup of the participant groups that use each device, rather than a consequence of differences in the devices alone. The full distribution of age-, education-, and gender-controlled test performance for each device group can be seen in Fig. 1. (see Table 4 for mean scores of each device group).

Study 2: Differences in reaction time by device, controlling for general cognitive ability

Study 1 indicated that there were differences in cognitive test performance between users of different devices. However, we also found differences in cognitive test performance on a test of vocabulary knowledge, where contributions of software/hardware characteristics should be minimal. Thus, it is possible that differences in reaction time measurement by device are related to differences in cognitive ability that vary with device ownership. In the next study, we recruited a group of participants to complete a larger set of timed tests and an untimed test of vocabulary knowledge, allowing us to examine how device differences differentially affect performance within the same participant, after controlling for device-insensitive differences in general cognitive ability (as measured by vocabulary). We also collected additional device information to be able to better distinguish between Android and Windows tablets and smartphones.

Methods

Participants

Participants consisted of 3818 volunteers who visited TestMyBrain.org over a period of 6 months in 2018. All tests completed by a single participant were completed over the course of a single session, ensuring that each set of tests associated with one participant were completed using a consistent device. Participants ranged in age from 18 to 80 years, with a mean age of 35.44 years (SD = 15.96). The

Table 4 Participant-level performance by device across tasks

Mean (SD) participant performance for frequently used device groups			
Device*	Simple RT (ms)	DSM** RT (ms)	Vocabulary accuracy (%)
Android	374.53 (75.13)	1494.44 (333.16)	63.79 (20.77)
iPhone	326.05 (56.34)	1418.30 (290.67)	67.23 (19.35)
iPad	346.94 (66.84)	1522.89 (335.70)	77.65 (17.78)
Macintosh	293.65 (48.97)	1342.29 (302.33)	77.81 (17.65)
Windows	304.44 (64.07)	1407.71 (377.70)	71.66 (21.16)
Effect size (95% CI) of device differences on test performance (Cohen's d, Windows device as comparison group, controlled for age, gender, and educational attainment)			
Android	1.12 (1.08, 1.16)	0.27 (0.21, 0.33)	−0.28 (−0.35, −0.21)
iPhone	0.44 (0.39, 0.48)	0.17 (0.10, 0.23)	−0.08 (−0.16, 0.00)
iPad	0.53 (0.47, 0.59)	−0.03 (−0.10, 0.04)	0.00 (−0.08, 0.07)
Macintosh	−0.11 (−0.15, −0.08)	−0.12 (−0.17, −0.08)	0.23 (0.19, 0.27)

*Android devices include both mobile phones and tablets; Windows devices include tablet, laptop, and desktop computers. For all measures, lower values indicate better performance

**Digit symbol matching test

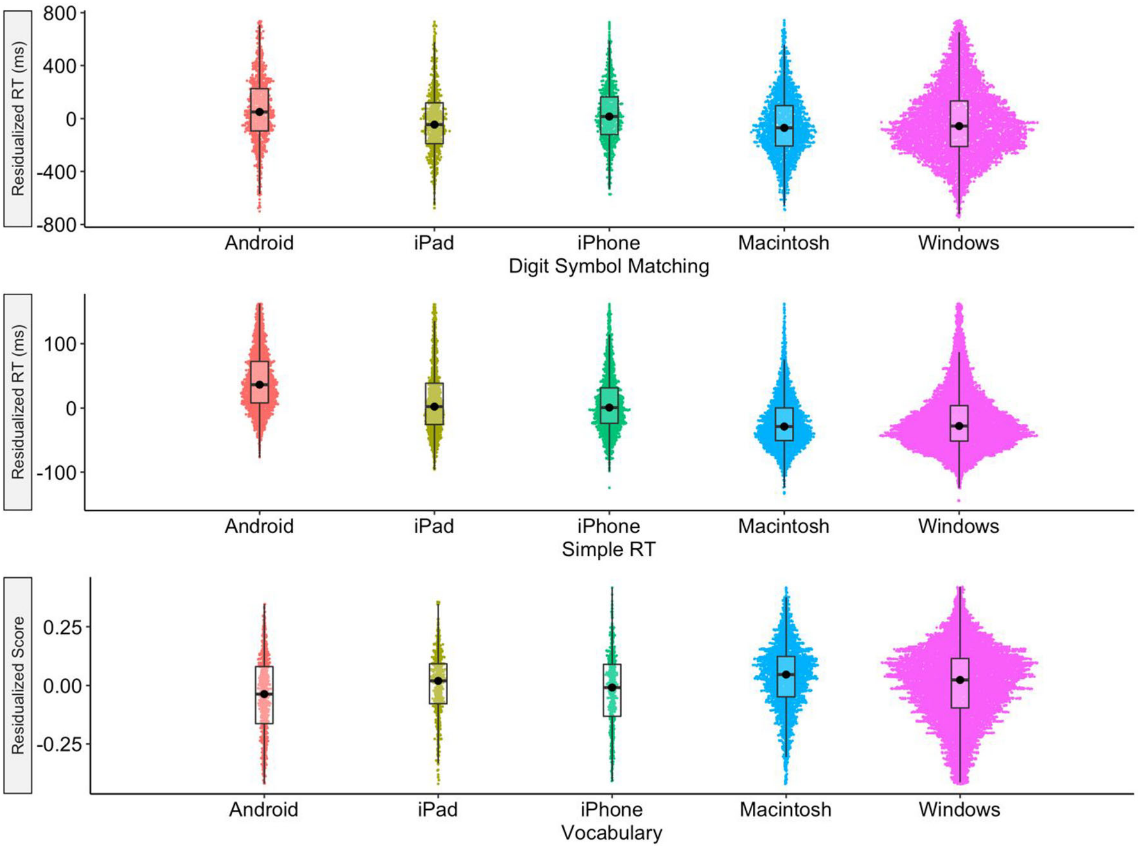

Fig. 1 Distribution of test performance by device

sample contained similar numbers of male and female participants (46.10% male, 51.70% female, 2.20% non-binary/genderqueer). While the TestMyBrain platform is open to participants in all parts of the world, we limited our sample to participants who reported speaking English as a native language. A total of 3515 participants completed the simple reaction time test, 3350 participants completed the digit symbol matching, 3259 completed the vocabulary test, 1997 completed part A of the trail-making task, and 1731 completed part B of the trail-making task (all tests have < 3818 completions due to partial battery completions in an online setting). Further information on the demographic characteristics of the sample can be found in Table 5.

Measures

For Study 2, participants completed a battery of tests including simple reaction time, digit symbol matching, and vocabulary. In addition, to look at potential effects of screen size on cognitive test performance, we included a digital trail-making task in a subset of later participants (*N*= 1997, parts A and B) (Bowie & Harvey, 2006; Reitan, 1958). The trail-making task, commonly used in neuropsychological assessment, requires participants to connect a series of points in a predetermined order as quickly as possible. In part A, the points are numbered and must be connected in numerical order (i.e. connecting point 1 to point 2, point 2 to point 3, etc.). In part B, some points are numbered while others are given letter labels, and participants must alternate numbered and letter-labeled points while maintaining numerical and alphabetical order (i.e. connecting point 1 to point A, point A to point 2, point 2 to point B, etc.). This task measures processing speed (parts A and B) and cognitive flexibility (part B). This test is scored based on the speed with which items are connected; the median time to connect two points in part A was 661.5 ms (*SD* = 264.35 ms), while in part B it was 1008 ms (*SD* = 462.97 ms).

Data analysis

All data analysis was performed in R, using the same packages used in Study 1 (R Core Team, 2012; Rudis et al., 2016; Torchiano, 2019). We examined differences in performance by device group using linear mixed-effects models. Reaction time on each trial was treated as a continuous dependent variable and was scaled to have a mean of 0 and a standard deviation of 1 for linear mixed-effects models, with results reported in standard deviation units. Participant ID and trial number were treated as random effects and device group, age (including quadratic and cubic effect of age), gender, age-

Table 5 Study 2 participant characteristics and analysis sample sizes

Variable [range]	Analysis *N*	*n* (%)	*M* (*SD*)
Age [18–80]	3818		35.44 (15.96)
Gender	3818		
Male		1760 (46.10)	
Female		1974 (51.70)	
Non-binary/genderqueer		84 (2.20)	
Education	2875		
Primary school		8 (0.28)	
Middle school		71 (2.51)	
High school		441 (15.61)	
Some college		711 (25.17)	
Technical training/associate's degree		206 (7.29)	
Bachelor's degree		782 (27.68)	
Graduate school		656 (23.22)	

gender interaction, and educational background were treated as fixed effects. Unlike in Study 1, we also included vocabulary performance as a fixed effect covariate to control for general cognitive ability, as this test is unlikely to be directly affected by device differences. Differences in vocabulary performance by device group were examined using a logistic mixed-effects model; accuracy on each trial was treated as a categorical (correct/incorrect) dependent variable, with participant ID and trial number treated as random effects and device group, age (including quadratic and cubic effect of age), gender, age-gender interaction, and educational background treated as fixed effects. In addition to these mixed-effects models, we also fit linear regression models for participant-level performance (median reaction time and vocabulary accuracy) with age, gender, educational background, and vocabulary performance as covariates; the effect size was computed based on residuals from these models using Cohen's *d*.

Outlying data were removed according to the same quality control criteria used in Study 1. A total of 129 participants (3.67%) had their simple reaction time data removed, while 93 participants (2.78%) had their digit symbol matching data removed, and two participants (0.05%) had their vocabulary data removed.

In this study, we also collected information on screen size and input type (mouse/keyboard or touchscreen) for all devices used by participants. Based on these data, we found that 13.68% of Windows devices were touchscreen-based tablets; the remainder were laptop or desktop computers.

Power

Data collection continued until we had at least 100 participants using each of the five most common device types (iPhone, Android phone or tablet, iPad, Macintosh laptop or desktop computer, or Windows laptop or desktop computer). Using a linear model to identify differences in performance between device groups, this sample size would provide 80% power to identify an effect size of f2 = 0.13 ($\alpha = 0.05$). Because some devices were more frequently used by participants, final group sizes ranged from 205 iPad users to 1623 Windows users; in addition, some participants did not complete the entire battery, so device groups vary in size between tests. Thus, our *a priori* power estimate should be considered a conservative estimation of the effect sizes detectable based on these data.

Results

Performance on timed tests after controlling for vocabulary

Consistent with findings from Study 1, mixed-effects linear regression modeling found significant differences between device groups (using Windows as a comparison group) after controlling for age, gender, and vocabulary in simple reaction time performance; reaction times on mobile devices were slower than reaction times on Windows devices (Android: $\beta = 0.58$, SE = 0.03, $t = 19.12$, $p < 0.001$; iPad: $\beta = 0.31$, SE = 0.05, $t = 6.74$, $p < 0.001$; iPhone: $\beta = 0.16$, SE = 0.03, $t = 5.25$, $p < 0.001$). There was no significant difference in reaction times for Macintosh laptop and desktop computers ($\beta = 0.01$, SE = 0.03, $t = 0.27$, $p = 0.79$). On the digit symbol matching task, we found that reaction times on Android devices were significantly slower than on Windows devices ($\beta = 0.06$, SE = 0.02, $t = 2.24$, $p = 0.03$), while reaction times on iPad devices were significantly faster ($\beta = -0.10$, SE = 0.04, $t = -2.61$, $p = 0.01$); other device groups did not show significant differences relative to Windows comparison group (iPhone: $\beta = 0.02$, SE = 0.03, $t = 0.71$, $p = 0.48$; Macintosh:

$\beta = -0.01$, SE = 0.02, $t = -0.53$, $p = 0.60$). Part A of the trail-making task showed slower reaction times among users of Android and iPhone devices, with no significant differences for other device groups (Android: $\beta = 0.18$, SE = 0.02, $t = 9.44$, $p < 0.001$; iPad: $\beta = -0.06$, SE = 0.03, $t = -1.92$, $p = 0.06$; iPhone: $\beta = 0.17$, SE = 0.02, $t = 8.40$, $p < 0.001$; Macintosh: $\beta = 0.02$, SE = 0.01, $t = -1.29$, $p = 0.20$); this was also true of part B (Android: $\beta = 0.16$, SE = 0.03, $t = 5.96$, $p < 0.001$; iPad: $\beta = -0.02$, SE = 0.04, $t = -0.64$, $p = 0.52$; iPhone: $\beta = 0.19$, SE = 0.03, $t = 7.06$, $p < 0.001$; Macintosh: $\beta = -0.01$, SE = 0.02, $t = -0.32$, $p = 0.75$). In general, users of mobile devices showed slower responses than users of laptop and desktop computers. Because vocabulary scores were included as covariates in these models, this indicates that differences in performance between device groups are unlikely to be due solely to differences in general cognitive ability between users of different devices, suggesting that differences in device characteristics are contributing to differences in measured performance. As in Study 1, device-related performance differences were larger for simple reaction time than those observed for digit symbol matching. Both trails A and trails B showed moderate effects for use of mobile phones compared to the Windows group (see Table 6).

Effects of screen size, input type, and browser on measured test performance

While device latency is one factor that varies between digital devices that can directly impact measured performance, other differences between digital devices may contribute to differences in observed performance. In addition to the hardware and software that contribute to differences in latency, digital devices differ in screen size, user interface, and precision with which one can select items on the screen. To examine the effects of these mechanisms, in further analysis, rather than classifying devices into categories based on hardware and OS, we examined differences in participant performance based on two device characteristics: screen size and input type. To determine the effects of screen size, we compared age-, gender-, and vocabulary-controlled performance of iPhone and iPad users; these devices use the same operating system, are made by the same company, and are both touchscreen-operated, but vary in size. To examine the effects of input type, we compared Macintosh laptop and desktop computers, which primarily use mouse and keyboard input, and iPad tablets, which primarily use touchscreen input; both devices are developed by the same company and have relatively large screens. By classifying devices based on specific characteristics, we can determine which differences between devices contribute most to differences in observed participant performance by observing whether one classification criterion shows greater effects for a particular test.

Using mixed-effects models controlling for age, gender, and educational attainment, with session ID and trial number included as random effects, we found a significant effect of screen size (comparing iPad and iPhone devices, with iPad as the comparison group) on the digit symbol matching test ($\beta = 0.16$, SE = 0.005, $t = 3.34$, $p < 0.001$) and parts A and B of the trail-making test (part A: $\beta = 0.18$, SE = 0.04, $t = 4.00$, $p < 0.001$; part B: $\beta = 0.15$, SE = 0.07, $t = 2.17$, $p = 0.031$); reaction times were slower on iPhones compared to iPads. There was no significant effect of screen size on performance on vocabulary ($\beta = -0.05$, SE = 0.12, $z = -0.40$, $p = 0.69$) or simple reaction time ($\beta = -0.06$, SE = 0.06, $t = -1.11$, $p = 0.27$). The effect of screen size on test performance was greatest on the trail-making tasks, with iPhone users showing slower performance than iPad users. This is consistent with our expectations based on the nature of these tests. Because these tasks require participants to connect targets without accidentally crossing over non-target points, clustering the points on a small screen makes it easier to unintentionally hit an incorrect point due to lack of precision, making the test more difficult. The small screen could also make it difficult to correctly identify the points to be connected. Effects on performance on other tests, both timed and untimed, were small or negligible. For effect sizes of screen size on all tests, see Table 7.

We also found a significant effect of input type (mouse vs. touchscreen) on test performance (comparing iPad and Macintosh devices, with iPad as the comparison group) on

Table 6 Age-, gender-, and vocabulary-controlled differences in performance by device (mean standard residual)

Effect size of device differences on age-, gender-, and vocabulary-controlled residualized participant performance (Cohen's d, Windows device as comparison group)

Device*	Simple RT	DSM RT	Trails A	Trails B
Android	1.10 (0.99, 1.22)	0.18 (0.07, 0.29)	0.50 (0.32, 0.67)	0.63 (0.46, 0.81)
iPhone	0.38 (0.27, 0.49)	0.09 (−0.02, 0.21)	0.45 (0.28, 0.63)	0.64 (0.46, 0.82)
iPad	0.70 (0.53, 0.86)	−0.14 (−0.30, −0.02)	−0.76 (−1.01, −0.51)	−0.09 (−0.34., 0.15)
Macintosh	0.05 (−0.04, 0.15)	−0.01 (−0.10, 0.09)	0.03 (−0.10, 0.15)	−0.01 (−0.13, 0.11)

*Android devices include both mobile phones and tablets; Windows devices include tablet, laptop, and desktop computers simple reaction time (β = −0.36, SE = 0.05, df = 701.98, t = −7.67, p < 0.001) and vocabulary (β = 0.26, SE = 0.11, z = 2.39, p = 0.02). There was no significant difference between these groups on digit symbol matching (β = 0.05, SE = 0.04, df = 741.7, t = 1.13, p = 0.26), trails A (β = 0.01, SE = 0.03, df = 472.0, t = 0.38, p = 0.71), or trails B (β = 0.01, SE = 0.04, df = 472.0, t = 0.37, p = 0.71). The effect of input method on simple reaction time performance was moderate, with users of mouse-operated devices showing faster performance than users of touchscreen devices. Users of iPad devices showed lower accuracy on the vocabulary assessment than users of Macintosh devices, though this effect was small. No other tests showed significant effects of input type. Input type is closely related to device latency; touchscreens typically have latency between 50 and 200 ms, while mouse input has a measured latency of approximately 20 ms (Henze et al., 2016; Pavlovych & Gutwin, 2012). Thus, the greater effect of input type (relative to screen size) on performance in simple reaction time suggests that the slower performance seen in users of smartphones is likely due to latency, rather than to screen size or other characteristics of the device. However, the significant effect on vocabulary performance suggests that there may be other participant-level differences between the device groups chosen that were not addressed by controlling for age, gender, and educational attainment.

To examine effects of browser differences on test performance, we compared performance between users of Safari, Google Chrome, and Firefox browsers, the three most commonly used browser types in our sample (using only data from participants using the main laptop/desktop-based versions of these browsers, rather than mobile versions). Controlling for age, gender, educational attainment, and vocabulary performance, linear mixed-effects models (using Safari as the comparison group) did not show any significant differences between browser groups on simple reaction time (Chrome: β = 0.07, SE = 0.04, t = 1.86, p = 0.06; Firefox: β = 0.01, SE = 0.05, t = 0.20, p = 0.84), digit symbol matching (Chrome: β = 0.03, SE = 0.03, t = 0.92, p = 0.36; Firefox: β = 0.01, SE = 0.04, t = 0.30, p = 0.77), trails A (Chrome: β = 0.03, SE = 0.02, t = 1.76, p = 0.08; Firefox: β = 0.04, SE = 0.02, t = 1.96, p = 0.05), or trails B (Chrome: β = 0.01, SE = 0.02, t = 0.63, p = 0.53; Firefox: β = 0.02, SE = 0.03, t = 0.82, p = 0.41). However, a mixed-effects logistic model showed a significant relationship between browser use and vocabulary accuracy, with use of Firefox devices associated with correct responses (Chrome: β = −0.04, SE = 0.08, z = −0.44, p = 0.66; Firefox: β = 0.23, SE = 0.11, z = 2.03, p = −0.04). The lack of variation on timed tests suggests that browser differences do not contribute significantly to differences in measured reaction time. However, the slight difference between browser groups' vocabulary performance suggests that there may be differences in participant characteristics between users of different browsers.

Discussion

In this manuscript, we looked at differences in cognitive test performance by device type, with a focus on the measurement of reaction times across different devices. We found three potential sources of variability in cognitive test performance related to device type. First, device ownership was related to differences in demographic characteristics associated with differences in reaction time. Second, device type was associated with differences in performance on an untimed measure (vocabulary), suggesting that cognitive differences between groups exist beyond what can be explained by age, gender, and education. Finally, after controlling for demographic characteristics and general cognitive ability, differences in device group were associated with differences in reaction time that were associated with operating system, input type (touch vs. mouse), and screen size. In general, measures that elicit shorter reaction times were most affected by factors that influence response time latency (operating system, mouse) whereas measures with larger, more complex stimuli and responses (trails A/B) were most affected by screen size. Our results suggest that differences in cognitive test performance between device groups are due to both the effects of the devices themselves and differences in the traits of the users who choose each device, beyond what can be explained by age, gender, and education.

To determine possible mechanisms contributing to these differences between device groups, we examined the effects of screen size and device input type on measured test performance. We found that the use of touchscreens had a stronger association with simple reaction time, while screen size had a

Table 7 Relationship between device characteristics and performance (age-, gender-, education-, and vocabulary-controlled standard residuals)

Test	Cohen's d (95% CI) (iPad vs. iPhone)	Cohen's d (95% CI) (iPad vs. Mac OS)
Simple reaction time	0.36 (0.17, 0.54)	0.72 (0.54, 0.90)
Digit symbol matching	−0.25 (−0.43, −0.07)	−0.14 (−0.31, 0.03)
Trails A	−0.80 (−1.09, −0.51)	−0.84 (−1.09, 0.58)
Trails B	−0.60 (−0.88, −0.31)	−0.09 (−0.34, 0.16)
Vocabulary (age-, gender-, and education-controlled)	−0.03 (−0.21, 0.15)	−0.33 (−0.50, −0.16)

stronger association with tests that require fine motor input and clear perception of the full screen; however, a small but significant association between use of non-touchscreen devices and vocabulary suggests that there may be differences between device user groups that go beyond age, gender, and educational attainment. Android touchscreen devices showed particularly strong associations with reaction time compared to iOS touchscreens. These differences were significant, with moderate to large effect sizes even after controlling for cognitive performance (as measured by vocabulary) as well as demographic variables.

The present study does have some important limitations. Although we controlled for demographic factors and performance on an untimed measure of cognitive ability, it is possible that there are other factors that differ between users of different devices. We also did not examine all possible differences between devices; in particular, it is possible that the iPad and iPhone devices differ in aspects other than screen size despite their shared manufacturer, operating system, and input type. Our analysis of the effects of screen size based on these device groups could thus be influenced by other factors. In addition, because digital devices are constantly changing with the introduction of new hardware and software, the specific differences in measured reaction time associated with each device may not be directly applicable to future versions of these devices. The change in digital devices over time may also introduce some heterogeneity into our device categories in study 1; because the data for that study were collected over a period of 5 years, it is possible that hardware and software changes over this period affected the effects of device characteristics on measured reaction time. However, the methods used here to determine the effects of device variability from the devices most commonly used in our sample could be applied to other devices used for cognitive testing. This study also did not measure differences in participants' internet connection speed; because all images in all tests used in this study are preloaded before the test begins, connection speed would not be expected to affect performance in this study, but this could be a factor in other web-based cognitive tests. Finally, our analysis of differences between major device types is dependent on the accurate recording and parsing of user agent strings; these strings are typically an accurate representation of the device being used to access a web page, but in a minority of cases they can provide inaccurate information on the device's characteristics (Kline et al., 2017).

Despite the widespread use of online cognitive testing tools in psychology research, this is the first study to directly examine the new sources of variability that mobile- and web-based research introduce in large samples of participants using a bring-your-own-device model (Germine et al., 2019). Previous research has examined the effects of device latency in cognitive testing and found significant device effects on the measurement of reaction times (Damian, 2010; Pronk et al., 2019). Our study expands on these findings by measuring the contribution of device variability to the measurement of reaction time in a naturalistic setting that draws on the full range of commercially available digital devices. Our online testing platform enabled us to recruit a large and demographically diverse sample of participants, which allowed us to examine the contribution of demographic factors to device group differences. The use of data from participants completing tests on their own devices, in a naturalistic setting, allowed us to go beyond existing measures of device variability, which typically focus on one aspect of a digital device, such as user interface or latency. While device differences introduce new sources of variability into cognitive testing, this variability can be measured and quantified, unlike many sources of variability in in-person testing such as measurement error and tester bias. If the variability introduced by these new tools is measured and appropriately mitigated, it could ultimately lead to greater precision in cognition research.

Mitigating the effects of device variability

While some of the effects of device variability can be mitigated simply by including device characteristics as covariates in analysis of data collected through web- and mobile-based assessments, there are also ways to further reduce this variance through study design. Users of smartphones typically showed the slowest reaction times on time-based tasks, indicating that these devices may produce artificially longer reaction times. Mobile phones can also cause problems on some tasks due to the small size of their screens and the difficulty of precise input. Thus, if it is important for the purposes of a study that reaction times be measured with minimal latency, one could choose not to make the test compatible with mobile phones and touchscreen devices and require participants to use a laptop or desktop computer. However, since many people access the internet regularly through their phones, this may result in lower participation. Furthermore, some study designs, such as ecological momentary assessment, require that participants be able to complete assessments on portable devices. There is a trade-off between the accuracy of reaction time-based assessments and the flexibility with which these assessments can be performed.

When participants are using a wide variety of digital devices, tests based on reaction time measurement may be less precise than tests based on other outcome measures, such as accuracy or biases in response. Although not all domains of cognitive function can be measured in this way, such measures avoid many of the sources of unquantified variation introduced by device variability. When reaction times are necessary, one way to reduce the effects of device latency on measures of participant performance is to choose tasks that produce longer response times. The efficacy of this approach can be seen by comparing the effects of device differences on performance in simple reaction time and performance in digit symbol matching. There are significant relationships for both tasks, but the effect size of input type is moderate to large for simple reaction time and relatively small for digit symbol matching. While both the digit symbol matching task and the simple reaction time task measure processing speed and are scored based on measured reaction time, digit symbol matching produces longer reaction times due to the additional cognitive challenge of matching each number to the appropriate image. Because the scale on which responses occur is greater, the proportion of variability attributable to the person—and presumably differences in processing speed rather than device effects—is correspondingly greater. Therefore, by choosing tasks that produce longer true response times, the proportion of observed variation that is due to device latency can be reduced, even if the latency itself remains the same.

Another way to mitigate the effects of device latency is to measure reaction times in ways that keep latency as consistent as possible, allowing reaction times to be more easily accounted for. For example, in tasks with multiple trials, one could measure reaction time as the time between when a response was registered for the previous trial and when a response is registered for the current trial. Since device latency is typically consistent within a single device over a short period of time, one can assume that the effect of device latency will be consistent from trial to trial. Thus, the time between registered responses is the same as the time between actual responses, even if all measurements are delayed by a consistent length of time reflecting device latency. Using this method of measurement, only the measured time between the start of the task and the first response will be altered by device latency; if one removes the first trial from analysis, the remaining measurements would be unaffected by latency. However, when tasks are presented this way, the time used to redirect attention from one trial to the next is included in the measured reaction time, which may not be desirable for all tasks. This approach also does not reduce the impact of user interface (separate from its relationship to latency) and screen size on performance.

While the previous solutions can minimize the effects of differences in device latency, different approaches are needed to avoid bias introduced by differences in user interface and screen size. In order to scale across devices of different sizes, stimuli must be simple enough to be easily viewable on small screens at the level of detail necessary to complete the task. Stimuli that are visually distinct and memorable on a laptop or desktop device may be difficult to differentiate on a mobile phone screen. The means by which a participant provides a response—pressing a key, clicking on a button, moving a cursor—must also be considered to ensure that differences between devices do not change the difficulty of the task. For example, if response buttons are placed close together (as in the trail-making tasks used in this study), it may be difficult to reliably select only the correct one when using a small touchscreen. The use of simple, scalable stimuli could help mitigate these effects in future research.

Each of these approaches can reduce the impact of device-related variability on measured performance on cognitive tests. However, there is no single approach that is appropriate for all forms of online research; each solution imposes some limitations on the research methods that can be used and the circumstances under which research can be conducted. When conducting online research in cognition, researchers will need to consider whether device variability could have a significant impact on their results, and if so, which approach is most appropriate to mitigate it. Whatever approach is chosen, simply measuring device characteristics in web- and mobile-based assessment is vital in any research in which latency, screen size, user interface, or other device characteristics could affect performance; without this information, the research methods made available by new technologies cannot be used to their full potential.

Open practices statement The datasets generated and/or analyzed during the current study are available in the Open Science Framework repository, https://osf.io/xf5v8/?view_only=8fedb487110e47198fc72f3d26bf5ccf. This research was not preregistered.

Funding This project was supported by funding provided by the National Institutes of Health to Laura Germine, including R01 DK121240 and R01 MH121617.

Declarations

Conflicts of Interest Laura Germine is a member of the scientific advisory board of Sage Bionetworks, a 501c3 nonprofit organization that distributes open source research tools, for which she receives a small honorarium. Dr. Germine is also Director of the Many Brains Project, a 501c3 nonprofit organization that develops open source cognitive assessments. She receives no financial compensation for this role.

References

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols, 1*(5), 2277-2281. https://doi.org/10.1038/nprot.2006.390

Buchanan, T. (2007). Personality testing on the Internet: What we know, and what we do not. In A. N. Joinson, K. Y. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 447–460). New York, NY: Oxford University Press.

Chaytor, N.S., Barbosa-Leiker, C., Germine, L.T., Fonseca, L.M., McPherson, S.M., Tuttle, K.R. (2020) Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. *The Clinical Neuropsychologist, 35*, 148–164.

Crawford, J. R. (1989). Estimation of premorbid intelligence: a review of recent developments. In J. R. Crawford & D. M. Parker (Eds.), *Developments in Clinical and Experimental Neuropsychology* (pp. 55-74). Springer US.

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards?

Behavior Research Methods, 42(1), 205-211. https://doi.org/10.3758/brm.42.1.205

de Bruijne, M., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: an experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review, 31*(4), 482-504. https://doi.org/10.1177/0894439313483976

de Liaño, B. G.-G., León, O. G., & Pascual-Ezama, D. (2012). Research participation improves student's exam performance. *The Spanish journal of psychology, 15*(2), 544-550. https://doi.org/10.5209/rev_SJOP.2012.v15.n2.38865

Deary, I. J., & Der, G. (2005). Reaction Time Explains IQ's Association with Death. *Psychological Science, 16*(1), 64–69. https://doi.org/10.1111/j.0956-7976.2005.00781.x.

Der, G., & Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. *Psychology and aging, 21*(1), 62-73. https://doi.org/10.1037/0882-7974.21.1.62

Drozdick, L. W., Wahlstrom, D., Zhu, J., & Weiss, L. G. (2012). *The Wechsler Adult Intelligence Scale—Fourth Edition and the Wechsler Memory Scale—Fourth Edition.* In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 197–223). The Guilford Press.

Dykiert, D., Der, G. , Starr, J. M. & Deary, I. J. (2012). Sex differences in reaction time mean and intraindividual variability across the life span. *Developmental Psychology, 48*(5), 1262–1276. https://doi.org/10.1037/a0027550.

Evarts, E. V., Teräväinen, H., & Calne, D. B. (1981). Reaction time in Parkinson's disease. *Brain : a Journal of Neurology, 104*(Pt 1), 167-186. https://doi.org/10.1093/brain/104.1.167

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

Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist, 33*(2), 271-286.

Germine, L. T., Garrido, L., Bruce, L., & Hooker, C. (2011). Social anhedonia is associated with neural abnormalities during face emotion processing. *NeuroImage, 58*(3), 935-945. https://doi.org/10.1016/j.neuroimage.2011.06.059

Gorus, E., De Raedt, R., Lambert, M., Lemper, J.-C., & Mets, T. (2008). Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology, 21*(3), 204-218. https://doi.org/10.1177/0891988708320973

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*(1), 877-902. https://doi.org/10.1146/annurev-psych-010814-015321

Greenwald, A.G., McGhee, D., Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1480.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D.Q., & Germine, L. T. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceeding of the National Academy of Sciences, 109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, *26*(4), 433–443. https://doi.org/10.1177/0956797614567339

Henze, N., Funk, M., & Shirazi, A. S. (2016). *Software-reduced touchscreen latency*. Paper presented at the Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, Florence, Italy.

Hetherington, C. R., Stuss, D. T., & Finlayson, M. A. J. (1996). Reaction time and variability 5 and 10 years after traumatic brain injury. *Brain Injury, 10*(7), 473-486. https://doi.org/10.1080/026990596124197

Hultsch, D. F., MacDonald, S. W. S., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *The Journals of Gerontology: Series B, 57*(2), P101-P115. https://doi.org/10.1093/geronb/57.2.P101

Jain, A., Bansal, R., Kumar, A., & Singh, K. D. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied & Basic Medical Research, 5*(2), 124-127. https://doi.org/10.4103/2229-516X.157168

Joy, S., Kaplan, E., & Fein, D. (2004). Speed and memory in the WAIS-III Digit Symbol—Coding subtest across the adult lifespan. *Archives of Clinical Neuropsychology, 19*(6), 759-767. https://doi.org/10.1016/j.acn.2003.09.009

Kline, J., Barford, P., Cahn, A., & Sommers, J. (2017). *On the structure and characteristics of user agent string*. Paper presented at the Proceedings of the 2017 Internet Measurement Conference, London, United Kingdom. https://doi.org/10.1145/3131365.3131406

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: report of Board of Scientific Affairs' Advisory Group on the conduct of research on the internet. *American Psychologist, 59*(2), 105-117. https://doi.org/10.1037/0003-066X.59.2.105

Luciano, M., Wright, M. J., Smith, G. A., Geffen, G. M., Geffen, L. B., & Martin, N. G. (2003). Genetic covariance between processing speed and IQ. In *Behavioral genetics in the postgenomic era.* (pp. 163-181). American Psychological Association.

Luciano, M., Wright, M. J., Geffen, G. M. et al. (2004). A Genetic Investigation of the Covariation Among Inspection Time, Choice Reaction Time, and IQ Subtest Scores. Behav Genet 34, 41–50. https://doi.org/10.1023/B:BEGE.0000009475.35287.9d.

McLean, S. A., Ressler, K., Koenen, K. C., Neylan, T., Germine, L., Jovanovic, T., ... Kessler, R. (2019). The AURORA Study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0581-3

Ngan, E. T. C., & Liddle, P. F. (2000). Reaction time, symptom profiles and course of illness in schizophrenia. *Schizophrenia Research, 46*(2), 195-201. https://doi.org/10.1016/S0920-9964(00)00027-X

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-Research: ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues, 58*(1), 161-176. https://doi.org/10.1111/1540-4560.00254

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166–180.

Passell, E., Dillon, D. G., Baker, J. T., Vogel, S. C., Scheuer, L. S., Mirin, N. L., et al. (2019). *Digital Cognitive Assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) Field Test Battery Report.* National Institute of Mental Health Report. Available online at: psyarxiv.com.

Pavlovych, A., & Gutwin, C. (2012). *Assessing target acquisition and tracking performance for complex moving targets in the presence of latency and jitter*. Paper presented at the Proceedings of Graphics Interface 2012, Toronto, Ontario, Canada.

Pronk, T., Wiers, R. W., Molenkamp, B., & Murre, J. (2019). Mental Chronometry in the Pocket? Timing Accuracy of Web Applications on Touchscreen and Keyboard Devices. *bioRxiv*, 639351. https://doi.org/10.1101/639351

R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Accessed March 2021.

Reips, U.-D. (2000). Chapter 4 - The web experiment method: advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 89-117). Academic Press.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills, 8*(3), 271-276. https://doi.org/10.2466/pms.1958.8.3.271

Rudis, B., Simon, L., & Langel, T. (2016). uaparserjs: Parse Browser 'User-Agent' Strings into Data Frames. R package version 0.1.0. https://CRAN.R-project.org/package=uaparserjs. Accessed March 2021.

Torchiano, M. (2019). *effsize: Efficient Effect Size Computation*. 10.5281/zenodo.1480624 (10.5281/zenodo.1480624), R package version 0.7.6, https://CRAN.R-project.org/package=effsize. Accessed March 2021.

Wilkinson, R. T., & Allison, S. (1989). Age and simple reaction time: decade differences for 5,325 subjects. *Journal of gerontology, 44*(2), P29-35. https://doi.org/10.1093/geronj/44.2.p29

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ, 3*, e1058. https://doi.org/10.7717/peerj.1058

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.