BRIEF REPORT

Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments

**Laura Germine · Ken Nakayama · Bradley C. Duchaine ·
Christopher F. Chabris · Garga Chatterjee ·
Jeremy B. Wilmer**

Abstract With the increasing sophistication and ubiquity of the Internet, behavioral research is on the cusp of a revolution that will do for population sampling what the computer did for stimulus control and measurement. It remains a common assumption, however, that data from self-selected Web samples must involve a trade-off between participant numbers and data quality. Concerns about data quality are heightened for performance-based cognitive and perceptual measures, particularly those that are timed or that involve complex stimuli. In experiments run with uncompensated, anonymous participants whose motivation for participation is unknown, reduced conscientiousness or lack of focus could produce results that would be difficult to interpret due to decreased overall performance, increased variability of performance, or increased measurement noise. Here, we addressed the question of data quality across a range of cognitive and perceptual tests. For three key performance metrics—mean performance, performance variance, and internal reliability—the results from self-selected Web samples did not differ systematically from those obtained from traditionally recruited and/or lab-tested samples. These findings demonstrate that collecting data from uncompensated, anonymous, unsupervised, self-selected participants need not reduce data quality, even for demanding cognitive and perceptual experiments.

Keywords Web-based testing · Cognition · Visual perception · Face recognition

L. Germine (✉)
Department of Psychology, Harvard University,
768 William James Hall, 33 Kirkland Street,
Cambridge, MA 02138, USA
e-mail: lgermine@gmail.com

K. Nakayama
Department of Psychology, Harvard University,
770 William James Hall, 33 Kirkland Street,
Cambridge, MA 02138, USA

L. Germine · J. B. Wilmer
Psychology Department, Wellesley College,
106 Central Street,
Wellesley, MA 02481, USA

B. C. Duchaine
Department of Psychological and Brain Sciences,
Dartmouth College,
Moore Hall 6207,
Hanover, New Hampshire 03755, USA

C. F. Chabris
Department of Psychology, Union College,
807 Union Street,
Schenectady, NY 12308, USA

G. Chatterjee
Department of Psychology, Harvard University,
33 Kirkland Street,
Cambridge, MA 02138, USA

Introduction

The emergence of the Web as a medium for conducting behavioral experiments has led to unprecedented opportunities for collecting large and broadly generalizable data sets with minimal resource investment (Buhrmester, Kwang & Gosling, 2010; Reips, 2000, 2007). However, while the validity and reliability of Web-based data have been demonstrated for many questionnaires (Buchanan, 2007; Gosling, Vazire, Srivastava & John, 2004; Reips, 2007), the validity of timed, performance-based, and/or stimulus-controlled experiments collected via the Web is less well established.

The question of data quality is increasingly critical. Behavioral research has typically relied on highly selected and homogeneous samples that constrain generalizability (Henrich, Heine & Norenzayan, 2010). Self-selected Web samples, on the other hand, tend to be much more diverse in terms of age, education, socioeconomic status, and geographic location (Gosling et al., 2004). In addition, Web-

based testing allows researchers to locate and study individuals with rare characteristics (e.g., Duchaine, Germine & Nakayama, 2007; Germine, Cashdollar, Düzel, & Duchaine, 2011). Indeed, the last several years have seen a rise in Web-based experiments across fields that rely on specific sample characteristics, including psychopathology (Germine & Hooker, 2011; Kendler, Myers, Potter & Opalesky, 2009), genetic epidemiology (Kendler et al., 2009), and behavioral genetics (Haworth et al., 2007; Wilmer et al., 2010).

A major issue that has yet to be comprehensively addressed outside of the self-report/questionnaire literature is whether data collected from self-selected samples (especially those that are uncompensated, anonymous, and unsupervised) can be comparable to data gathered in lab-based settings. A lack of tangible incentives could result in careless responding, lack of focus, or even deception (Kraut et al., 2004). In addition, differences between participants in their levels of technical expertise, the computer systems that they use to access the Internet, and other variables could disproportionately impact measurement accuracy or reliability in self-selected Web samples relative to lab samples (Buchanan, 2007; Buchanan & Smith, 1999). As a result, researchers doing Web-based experiments can encounter skepticism from reviewers and editors regarding the reliability and validity of Web data (Gosling et al., 2004; Kraut et al., 2004), especially in disciplines in which Web-based methods have not been widely adopted.

Previous studies have produced qualitative evidence for convergent validity between Web- and lab-based cognitive/perceptual tests by showing that they tap into the same basic psychological constructs (Birnbaum, 2000; Krantz, Ballard & Scher, 1997; Krantz & Dalal, 2000; McGraw, Tew & Williams, 2000). This type of qualitative equivalence (Buchanan, 2007) has been demonstrated for a range of cognitive and perceptual tests, including tests involving very brief, tachistoscopic presentation times (McGraw et al., 2000), measurement of subtle variations in response times (McGraw et al., 2000; Nosek, Banaji & Greenwald, 2002), accurate perception of simple visual stimuli (Krantz et al., 1997; Krantz & Dalal, 2000; McGraw et al., 2000; Senior, Phillips, Barnes & David, 1999), and focused attention (Birnbaum, 2000; Krantz & Dalal, 2000; McGraw et al., 2000). Moreover, a growing number of large-scale Web-based studies of cognition and perception have now been published that have demonstrated reliable, replicable, and theoretically consistent findings (Germine, Duchaine, & Nakayama, 2011; Germine & Hooker, 2011; Nosek et al., 2002; Soto, John, Gosling & Potter, 2011).

While these qualitative findings support the core validity of cognitive and perceptual data collected via the Web, they do not clearly falsify the common notion that Web data are necessarily noisier than lab data (Gosling et al., 2004; Kraut et al., 2004). Such noise could result from numerous sources, including not only lack of diligence from participants, but also the influence of a variety of situational and personal

variables that tend to differ between Web and lab (Buchanan & Smith, 1999; Kraut et al., 2004). One method for assessing data quality for self-selected Web and lab samples is by comparing their performance means, variances, and measurement reliability (Buchanan, 2007). Poorer overall data quality from self-selected Web samples should be reflected in a systematic pattern of lower performance means, greater variance, and/or lower measurement reliability. Demonstrating comparability between Web and lab on these metrics for challenging, performance-based tests would provide a proof of concept that Web data need not be of lower quality than lab data for performance-based cognitive and perceptual experiments.

Here, we present data gathered using our website, TestMyBrain.org, for a range of challenging cognitive and perceptual tests. We compare the performance of these Web samples to the performance of samples tested and/or recruited in the lab in order to assess data quality.

General method

Participants

TestMyBrain.org is a Web-based testing environment that provides personalized performance feedback in exchange for study participation and has collected data from half a million participants over the past three years. The participants in this study were visitors to TestMyBrain.org from 2009 to 2011 who took part in experiments in exchange for feedback on how their performance compared with the performance of the average participant. TestMyBrain.org attracts participants of all ages and education levels (approximately 60 % female; average age 30 years, $SD = 13$). Most of the participants have some college education. Visitors to TestMyBrain.org come primarily from the United States (55 %) and from other English-speaking countries (21 %), with the remaining 24 % coming from almost every other country in the world. TestMyBrain.org receives traffic primarily through social networking sites and search engines. In the past three years, for example, the top traffic sources for TestMyBrain.org have been www.google.com, www.stumbleupon.com, and www.i-am-bored.com. The most popular search terms leading to TestMyBrain.org (accounting for 13 % of traffic) were "brain test" and "brain tests." This suggests that many visitors arrive at the website because they are curious about their cognitive abilities.

The data for the three studies reported here were taken from three different batteries of tests, included on TestMyBrain.org at different times. Table 1 shows the names of the experiment batteries, as advertised on the front page. Many experiments have been run on TestMyBrain.org since its creation in 2008. For the purposes of this study, however, we only selected those tests that (a) have been described and validated in peer-reviewed studies and (b) have good psychometric properties.

Table 1 Exclusion criteria for Web samples

Web Sample Exclusions	Technical Problems/Cheating	Repeat Participation	Age out-of range	Inappropriate device	Total Excluded
Words, Faces, and Abstract Art 4,427 participants completed *Includes VPAM, CFMT, AAM*	150(3.4%)	126(2.8%)	45(1%)	61(1.4%)	347(7.8%)
Keeping Things in Mind: 25,505 participants completed *Includes FDS*	1267(5%)	1470(5.8%)	415(1.6%)	496(1.9%)	3315(13%)
Reading Faces: 3,429 participants completed *Includes CFMT, RMIE*	99(2.9%)	71(2.1%)	51(1.5%)	36(1%)	238(6.9%)

The tests here were part of three different batteries posted on TestMyBrain.org at different times. The batteries were advertised on the front page using the names given in bold. Shown are the number of participants who completed each set of tests, the tests included in that set, and a breakdown of the participant numbers (and percentages of total completers) who were excluded from the final sample. The reasons for exclusion were as follows. (1) Technical problems/cheating: Participants indicated by ticking a box that they had technical problems or that they had used strategies that could be considered cheating. (2) Repeat participation: Participants answered "No" to the question "Is this your first time taking this test?" (3) Age out of range: Participants reported an age less than 10 or greater than 70 years old. (4) Inappropriate device: Participants indicated that they were using a device other than a desktop or laptop computer to complete the tests

The participants were people who completed all of the tests in each battery (Reips, 2002). A subset of the participants were then excluded from the analysis for self-reported technical problems, self-reported cheating, repeat participation, age less than 10 years or greater than 70 years, and use of a device that was not a desktop/laptop computer. The numbers and percentages for each of these types of exclusions are shown in Table 1.

Informed consent was obtained from all participants in accordance with the guidelines set by the Committee for the Use of Human Subjects at Harvard University and Wellesley College.

Cognitive/perceptual tests

The tests were chosen on the basis of their psychometric properties, sensitivity to individual differences in healthy samples, and ease of administration in an unsupervised setting. These include tests of face memory (the Cambridge Face Memory Test; CFMT), emotion perception (Reading the Mind in the Eyes; RMIE), visual memory (Abstract Art Memory Test; AAM), verbal episodic memory (Verbal Paired Associated Memory; VPAM), and working memory (Forward Digit Span; FDS).

All of the included tests are challenging enough to tap into a wide range of individual differences, with minimal ceiling or floor effects. Three of the five (CFMT, RMIE, and AAM) include complex visual stimuli and might be affected by any loss of stimulus control. Four of the five (CFMT, AAM, VPAM, and FDS) have limited presentation times, and so rely on focused participant attention. Two of the five (VPAM and FDS) depend on participant honesty and would be heavily susceptible to cheating.

All of the tests are administered with comprehensive instructions—and practice trials, where needed—and can be run without experimenter assistance.

Cambridge Face Memory Test (CFMT) The CFMT is a widely used test of unfamiliar face recognition in which participants are required to learn and then recognize six male target faces in novel poses and lighting (Duchaine & Nakayama, 2006). The CFMT consists of 72 test items, over three phases of increasing difficulty. The test starts with a brief, three-item practice, followed by an 18-item introductory/learning phase. In this phase, participants are shown three views of each target face for 3 s per view, followed by three-item forced choice trials in which the participants must distinguish the target face from two nontarget faces. Here, the target face images were identical to the images used at encoding. The moderate-difficulty phase follows, with 30 three-item forced choice trials using novel target images (see Fig. 1). The final, most difficult phase includes 24 items in which novel target images have the face features obscured with Gaussian noise.

Reading the Mind in the Eyes (RMIE) The RMIE is a widely used measure of complex face emotion processing (Baron-Cohen, Wheelwright, Hill, Raste & Plumb, 2001). In this test, participants see the eye region of a face and are asked to choose which of four complex emotion words best describes that emotion (see Fig. 1). The test includes 36 items in total, plus one practice item at the beginning.

Abstract Art Memory (AAM) The AAM is a test of visual recognition memory (Wilmer et al., 2010). In this test, participants are shown 50 images of unfamiliar abstract art images, one at a time for 1 s each. Afterward, participants are given 50

Fig. 1 Sample stimuli from our perceptual tests. Examples of trials are shown from three of the five tests included in Studies 1–3, but with images different from those included in the actual tests

Cambridge Face Memory Test (CFMT)
Learning phase   **Testing phase**

Reading the Mind in the Eyes (RMIE)
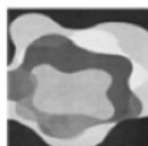curious   excited

suspicious   terrified

Abstract Art Memory (AAM)
Learning phase   **Testing phase**

three-item forced choice trials in which they are asked to distinguish target images from nontarget images (see Fig. 1). Three practice items are presented at the beginning of the test.

Verbal Paired Associates Memory (VPAM) The VPAM measures verbal episodic memory, requiring participants to learn and then recognize specific word pairs after a delay (Wilmer et al., 2010). During the memorization phase, participants are presented with 25 word pairs, presented on screen one at a time, for 6 s per word. In the test, participants are shown one of the words in each pair and are asked to choose the second word from a list of four words. Notably, the three distractor words are themselves part of other word pairs (Wilmer et al., 2010).

Forward Digit Span (FDS) The FDS test was adapted from the Wechsler Adult Intelligence Scales (Wechsler, 2008), but with visual presentation of the numbers. In this test, participants are shown sequences of digits on screen for 1 s per digit (e.g., a total of 3 s for three digits). Immediately after seeing each sequence, participants are asked to type in the sequence that they just saw. The test starts with two practice sequences. For test trials, sequences start at a length of two digits (two trials per sequence length) and continue to

an upper limit of 11 digits or until the participant gets both sequences of a given length incorrect.

Study 1: Variability across samples

We first investigated how the results from a Web sample compare with the normal variations seen across different labs for a challenging test of face perception and memory, the CFMT (Duchaine & Nakayama, 2006).

Method

Summary data from lab samples were taken from published studies in which means, standard deviations, and Cronbach's alpha reliability coefficients (Cronbach, 1951) were reported for large, healthy samples (Bowles et al., 2009; Duchaine & Nakayama, 2006; Herzmann, Danthiir, Schacht, Sommer & Wilhelm, 2008).

The participants for this study were those who took part in an experiment advertised as "Words, Faces, and Abstract Art." Table 2 gives the characteristics of the participants in each sample.

Table 2 Summary characteristics for Web and lab samples

Sample Characteristics	Sample Size		Age-M(SD)		% female	
Study 1: Variability Across Samples	web	lab	web	lab	web	lab
Web vs. CFMT Lab 1 (Herzmann)	4080	153	26(11)	24(5)	65	52
Web vs. CFMT Lab 2 (Bowles)	4080	124	26(11)	23(18-32 range)	65	59
Web vs. CFMT Lab 3 (Duchaine)	4080	50	26(11)	20(2)	65	58
Study 2: Sex and age-matched	web	lab	web	lab	web	lab
CFMT	447	45	19(2)	19(2)	100	100
RMIE	447	45	19(2)	19(2)	100	100
VPAM	971	45	19(2)	19(2)	100	100
FDS	1486	60	19(2)	19(2)	100	100
Study 3: Recruitment Method	web	ATR	web	ATR	web	ATR
CFMT	1028	578	37(11)	37(11)	61	62
AAM	1028	129	37(11)	38(11)	61	62
VPAM	1028	120	37(11)	38(11)	61	62
FDS	1561	759	42(15)	42(15)	62	62

"Web" refers to samples made up of self-selected, uncompensated, unsupervised, and anonymous participants who followed links from other websites or search engines to TestMyBrain.org. In Study 1, the same Web sample was compared with the lab samples from three published studies. In Studies 2 and 3, Web samples were age- and sex-matched to lab samples. In Study 2, the same Web sample was run on the CFMT and RMIE, and different Web samples were run on the VPAM and the FDS. In Study 3, "ATR" refers to traditionally recruited, nonanonymous participants who were enrolled in the study by the Australian Twin Registry and were then tested over the Internet. In Study 3, the same Web sample was run on the CFMT, AAM, and VPAM, and a different Web sample was run on the FDS test. The ATR samples were different for all four tests, but with significant overlap

All comparisons had a power of $\geq .95$ to detect medium effect sizes (Cohen's $d = 0.5$), and a power of .8 to detect effect sizes from 0.2 to 0.35 (Cohen, 1992).

Results

The CFMT results from a Web sample and from three lab samples are shown in Fig. 2a. The reliabilities and variances were similar for all comparisons, with a trend toward greater variance in the Web sample than in one lab sample (Duchaine & Nakayama, 2006) and greater measurement reliability in the Web sample than in another lab sample (Herzmann et al., 2008). Furthermore, there were no systematic mean differences in performance between the Web and lab samples: The Web sample had significantly higher scores than in one lab sample (Herzmann et al., 2008), comparable scores to those of a second lab sample (Bowles et al., 2009), and significantly lower scores than in a third lab sample (Duchaine & Nakayama, 2006). The results of statistical comparisons are given in Table 3. Cohen's d effect sizes (Cohen, 1992) and confidence intervals for each comparison are shown in Fig. 3a.

As the CFMT can be broken down into three phases of increasing difficulty, we also compared performance on each of these phases. All samples (Web and lab) showed reduced performance means in more difficult phases of the test. In the introductory learning phase, there were no significant differences in mean performance between the Web and lab samples.

In the more difficult second and third phases, the patterns of mean differences shown in Fig. 2 and Table 3 were replicated ($p < .05$, two-tailed).

According to these data, the results from our Web samples on the CFMT are comparable to those from a range of lab samples, with no systematic mean differences in performance (see Fig. 2a).

Study 2: Age- and sex-matched samples

Although age and sex effects are small for the CFMT, they are nonzero (Bowles et al., 2009; Germine, Duchaine, & Nakayama 2011). We therefore wanted to compare results in age- and sex-matched samples, as well as across a broader range of tests. The experiments in this study were advertised as "Words, Faces, and Abstract Art," "Keeping Things in Mind," and "Reading Faces." Information about the participant characteristics is shown in Table 2.

Method

For this second study, we analyzed data from the CFMT, RMIE, VPAM, and FDS tests.

The lab participants were female undergraduate students at a selective liberal arts college (Wellesley College), were
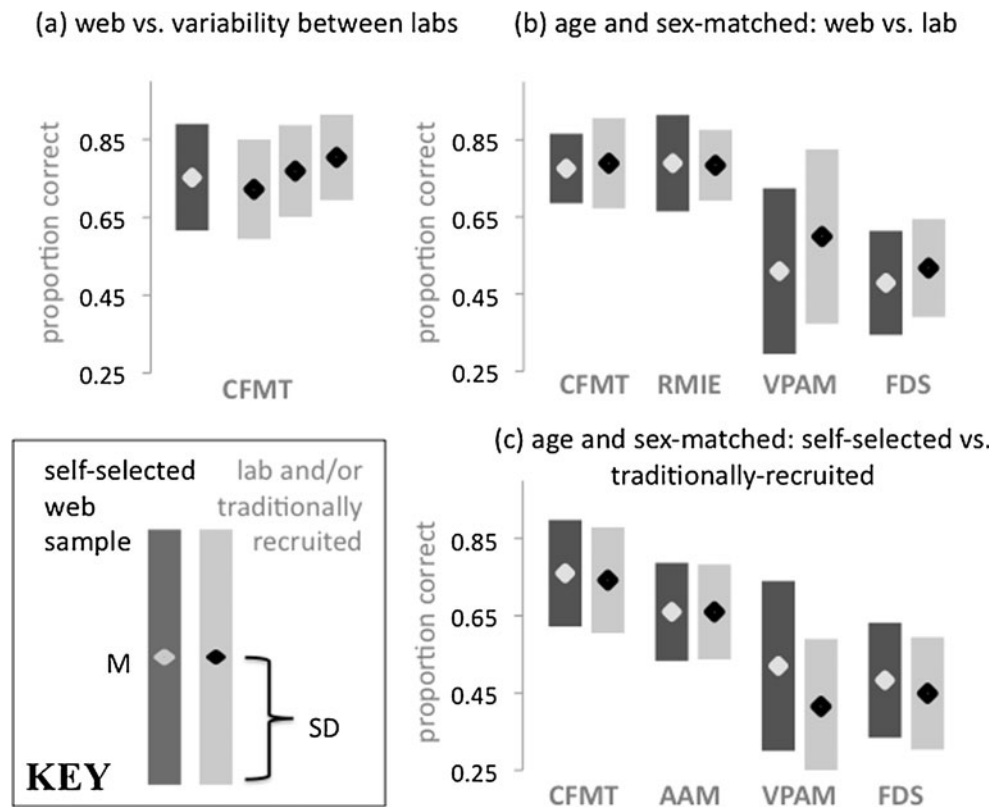
Fig. 2 Results for self-selected Web samples and traditional samples. Diamond shapes indicate mean performance levels for the self-selected Web samples (light diamonds in dark bars) and traditional/comparison samples (dark diamonds in light bars). The bars show ±1 standard deviation in performance for each measure. The tests included are the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), the Reading the Mind in the Eyes Test (RMIE; Baron-Cohen et al., 2001), the Verbal Paired Associates Memory Test (VPAM; Wilmer et al., 2010), the Abstract Art Memory Test (AAM; Wilmer et al., 2010), and the Forward Digit Span Test (FDS; adapted from Wechsler, 2008). (**a**) Results from a self-selected Web sample showed variance comparable to two of three published lab samples, with no systematic mean differences. (**b**) Results from a lab sample of female undergraduates

from ages 18 to 21 who spoke English as a native language were compared with self-selected age- and sex-matched Web samples who also spoke English as a native language. The variances were comparable across tests. Mean performance was comparable for two of the tests, but self-selected Web participants had lower mean performance on the IQ-loaded tests (VPAM and FDS). **c** Results from traditionally recruited, Web-tested samples were compared with self-selected age- and sex-matched Web samples. The variances were comparable across tests. Mean performance was comparable for one test, but self-selected Web participants had *higher* means on the three other tests, including two IQ-loaded tests (VPAM and FDS), in contrast with the pattern in Study 2 (panel b)

18–21 years of age, and spoke English as a native language. For comparison purposes, the Web samples were filtered to include only female participants from the ages of 18 to 21 who spoke English as a native language.

Importantly, in the FDS and VPAM tests, participants can cheat by writing down the sequences/word pairs. Both the VPAM and FDS scores are also expected to be positively correlated with IQ (Uttl, Graf & Richter, 2002; Wechsler, 2008). Wellesley students, on average, score at the 97th percentile on SAT reasoning test scores among female test-takers (College Board, 2012; Wellesley College, n.d.), indicating high-average IQs (Frey & Detterman, 2004). Thus, unless the Web participants were systematically cheating on these tests, we would expect poorer performance on the VPAM and FDS in our Web samples, as they are likely to be closer to the population mean in IQ than were our lab

samples. CFMT and RMIE scores, on the other hand, are not significantly related to IQ (for CFMT, see Bowles et al., 2009; for RMIE, see Baron-Cohen et al., 2001). Thus, mean performance differences between our two samples on the CFMT and RMIE would not be predicted.

All comparisons had a power of $\geq .9$ to detect medium effect sizes (Cohen's $d = 0.5$) and a power of .8 to detect effect sizes from 0.33 to 0.38 (Cohen, 1992).

Results

Results from the age- and sex-matched Web and lab samples are shown in Fig. 2b. Mean scores were similar on the CFMT and RMIE, but lower for the Web than for the lab samples on the VPAM and the FDS. Cohen's d effect sizes and confidence intervals are shown in Fig. 3b. Variance in

Table 3 Comparisons of means, variances, and internal reliabilities for self-selected Web and traditional samples

Main Comparisons	Comparison of Means			Comparison of Variances			Comparison of internal reliability			
Study 1: Variability Across Samples	t	df	p	F	df	p	a(web)	a(lab)	Z	p
Web vs. CFMT Lab 1 (Herzmann)	2.6	4231	0.01	1.03	4079, 152	0.41	0.9	0.86	2	0.05
Web vs. CFMT Lab 2 (Bowles)	1.6	4202	0.11	1.17	4079, 123	0.12	0.9	0.88	0.92	0.36
Web vs. CFMT Lab 3 (Duchaine)	2.9	4128	0.004	1.4	4079, 49	0.07	0.9	0.86	1.14	0.25
Study 2: Sex and age-matched	t	df	p	F	df	p	a(web)	a(lab)	Z	p
CFMT	0.01	490	0.99	1.14	446, 44	0.3	0.88	0.87	0.43	0.67
RMIE	0.57	490	0.57	0.96	446, 44	0.59	0.55	0.52	0.21	0.83
VPAM	2.7	971	0.007	0.9	927, 44	0.71	0.82	0.85	-0.64	0.52
FDS	1.9	1486	0.05	1.13	1442, 59	0.28	0.67*	0.68*	-0.14	0.89
Study 3: Recruitment Method	t	df	p	F	df	p	a(web)	a(ATR)	Z	p
CFMT	2.5	1604	0.01	1.01	1027, 577	0.42	0.9	0.89	0.87	0.38
AAM	0	1155	1	1.07	1027, 128	0.33	0.79	0.77	0.65	0.52
VPAM	7.7	1146	1E-04	1.58	1027, 119	9E-04	0.82	0.72	2.8	0.005
FDS	3.5	2318	4E-04	1.05	1560, 758	0.23	0.64*	0.64*	-0.26	0.79

Means for self-selected Web and traditional samples were compared using two-tailed, independent-samples *t* tests. Variances were compared using *F* tests (ratio of the squared standard deviations in each sample). Internal-consistency reliability was calculated using Cronbach's alpha for all tests except the Forward Digit Span test (FDS), for which Spearman–Brown's split-half reliability (indicated with an asterisk) was calculated by dividing the data between first and second trials for each sequence length. The reliability coefficients were compared using Fisher *r*-to-*Z* transformations. For each comparison, statistically significant differences are underlined

performance and the internal reliability measures were similar for all tests (see Table 3).

When the Web and lab samples were not matched (i.e., the Web samples included all ages, males, and nonnative English speakers), all comparisons yielded the same results as in the matched samples, except that the unmatched Web sample had larger standard deviations than did the lab sample in scores on the FDS ($p < .05$) and on the CFMT and RMIE (trend: $p = .07$ for both). Thus, matching the Web sample to the lab sample for age, sex, and native language impacted the variances in the Web sample.

As before, the CFMT could be broken into three phases of increasing difficulty. Like the lab sample, the Web sample showed reduced mean performance on each subsequent phase, reflecting increases in difficulty, with no significant differences in mean performance when the lab and Web samples were compared for each phase separately.

On the basis of these data, the results from our Web samples are comparable to those from an age- and sex-matched lab sample on most measures, with mean differences consistent with high performance in the Wellesley College sample on two IQ-loaded tests (VPAM and FDS). Notably, these mean differences are not consistent with widespread cheating among Web participants. CFMT and RMIE performance have no known relationship with IQ (Baron-Cohen et al., 2001; Bowles et al., 2009), and mean performance was similar on these tests in the Web and lab samples. Altogether, these results indicate comparable data quality in age-/sex-matched Web and lab samples.

Study 3: Self-selected versus traditionally recruited samples

In this study, we wanted to specifically address the issue of self-selection versus traditional selection in the quality of Web data. Unlike TestMyBrain.org volunteers, participants acquired through traditional methods are generally individually contacted/recruited and are often uniquely identifiable, and thus might take participation more seriously. In this study, we compared two groups of participants differing in their methods of recruitment/selection, but using the same Web-administered tests.

Method

For this comparison, we compared data from the CFMT, AAM, VPAM, and FDS. Privately recruited participants were members of the Australian Twin Registry (ATR), an organization that recruits and facilitates twin studies through a national registry of 31,000 Australian twins who had contributed to 655 peer-reviewed publications as of 2008 (ATR, n.d.). The ATR samples included here were part of a twin study of face recognition ability (Wilmer et al., 2010; CFMT, VPAM, and AAM) and of another, ongoing twin study (FDS). For the face recognition study, the ATR samples originally completed the CFMT, and then were invited back to complete the VPAM and AAM tests. Self-selected Web participants were normal visitors to TestMyBrain.org who clicked on links to participate in
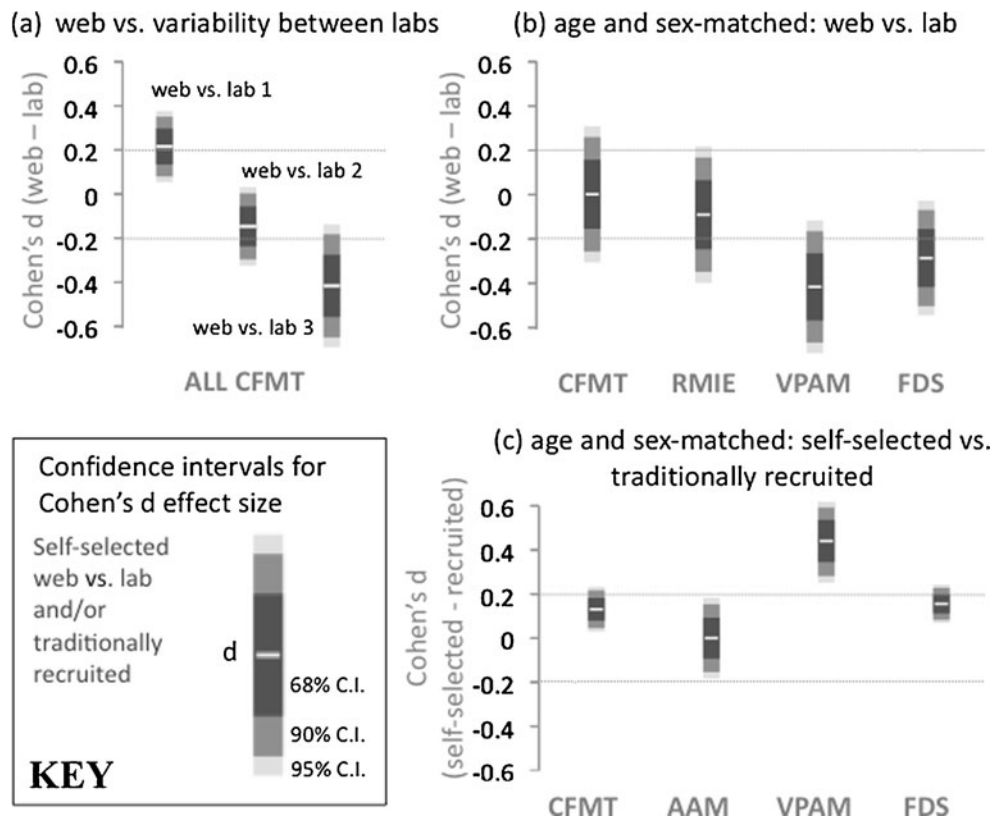
Fig. 3 Confidence intervals for effect size differences between self-selected Web samples and traditional samples. For each bar, the white dash shows the difference in mean performance between the Web and comparison samples in terms of Cohen's *d* (mean difference divided by pooled standard deviation; Cohen, 1992). The confidence intervals for Cohen's *d* are also shown, with the dark gray portion of the bar showing the 68 % confidence interval, the medium gray portion of the bar showing the 90 % confidence interval, and the light gray portion of the bar showing the 95 % confidence interval. Dashed lines are shown at Cohen's *d* values of –0.2 and 0.2, typically considered small effect sizes (Cohen, 1992). **a** Cohen's *d* confidence intervals are shown for a comparison of a Web sample versus three separate lab samples. The Web sample showed better performance than Lab 1 (Herzmann et al., 2008), performance comparable to that of Lab 2 (Bowles et al., 2009), and poorer performance than Lab 3 (Duchaine & Nakayama, 2006). **b** Cohen's *d* confidence intervals are shown for the comparison of Web samples to a lab sample across four tests. The Web sample was age- and sex-matched to the lab sample. **c** Cohen's *d* confidence intervals are shown for the comparison of self-selected Web samples to traditionally recruited samples across four tests. The Web samples were again age- and sex-matched to the traditionally recruited samples. Across studies, there was no systematic tendency toward lower scores in the self-selected Web sample as compared with lab and/or traditionally recruited samples. Differences in both directions were found mostly for the IQ-loaded tests

the "Words, Faces, and Abstract Art" and "Keeping Things in Mind" experiments. The self-selected Web samples completed the VPAM and AAM tests, followed by the CFMT. We matched Web samples to the ATR samples for age and sex by randomly sampling (without replacement) from the original Web sample to create a new Web sample that was similar to the ATR sample in both proportions of females and proportions of participants at each age from 18–65 (see Table 2).

Gestational factors are associated with a six-point childhood reduction in IQ scores in twins relative to nontwins (Ronalds, DeStavola & Leon, 2005) that persists in attenuated form into adulthood (Plomin, DeFries, McClearn & McGuffin, 2008).

Given that our ATR sample was composed entirely of twins, we predicted that the ATR twins might show reduced mean performance on the IQ-loaded tests (VPAM and FDS) relative to self-selected Web participants. We did not predict differences in the CFMT or AAM tests, as these are not known to vary with IQ.

All comparisons had a power of $\geq .99$ to detect medium effect sizes (Cohen's $d = 0.5$) and a power of .8 to detect effect sizes from 0.1 to 0.24 (Cohen, 1992).

Results

Mean performance was comparable for the AAM test but was significantly higher in our self-selected Web sample than in the ATR sample for all other tests (see Fig. 2c). The effect sizes for all comparisons are shown in Fig. 3c. The performance variances and internal reliability were similar for the two samples across all tests, with the exception of the VPAM, on which the self-selected Web data were significantly more reliable (see Table 3).

When the Web and ATR samples were not age- and sex-matched, all comparisons yielded the same results, except (1) there was no longer a significant difference in mean performance on the CFMT, and (2) the unmatched Web sample had larger standard deviations in scores on the FDS ($p < .05$).

In the CFMT, both the Web and ATR samples showed reduced mean performance on more difficult stages of the test, but the Web sample scored significantly higher on both the second and third phases of the test ($p < .05$, two-tailed), with no differences in the first, introductory phase.

Notably, the mean differences between the self-selected and ATR-recruited samples for the VPAM and FDS were in the opposite direction from the mean differences between the Web and lab samples in Study 2. This likely reflects slightly lower IQ in our twin ATR sample than in our self-selected Web-based sample and is not consistent with poorer data quality for self-selected Web-based samples.

Discussion

Here we have shown that for challenging timed and/or stimulus-controlled cognitive and perceptual experiments, self-selected Web samples can yield data that are comparable in quality to data collected in the lab in terms of performance measures, even when those self-selected Web participants are anonymous, uncompensated, and unsupervised. Our results indicate that data from self-selected Web samples need not involve a trade-off between participant numbers and data quality.

We chose performance-based tests in cognition and perception that have good psychometric properties and capture a wide range of individual differences. These tests also depend on accurate timing, sustained participant attention, participant honesty, and accurate perception of complex stimuli. Reduced focus or conscientiousness in self-selected Web volunteers should result in systematic differences in test means, variances, and/or measurement reliability. Although the results were comparable between Web and lab for most tests and measures, differences in mean performance were observed for the IQ-loaded tests. These differences were not systematic, however: Our self-selected Web sample performed better in some cases (Studies 1 and 3) and more poorly in others (Studies 1 and 2), and the Web data were as reliable as the lab data in all comparisons. The fact that performance differences were found in both directions with no reduction in reliability suggests that these differences were not related to overall data quality but to characteristics of the comparison sample (Krantz & Dalal, 2000).

In the late '90s, pioneering work by early Web experimenters showed that cognitive and perceptual experiments on the Web tended to replicate experimental effects observed in the lab (Birnbaum, 2000; Krantz et al., 1997; Krantz & Dalal, 2000; McGraw et al., 2000). Between the time that these early studies were run and the data in the present study were collected, however, Internet adoption has increased from 36 % to 80 % in the United States (1999–2009), with greater usage for social or recreational purposes (Pew Internet Research, 2012a, b). Given these changes, the present study provides a contribution to the literature by addressing data quality through performance metrics, but also by providing a contemporary look at Web data, given the modern Internet landscape of social and recreational users from a broad range of backgrounds.

Although we found no evidence of systematic differences between self-selected Web samples and traditional samples, the degree to which self-selected Web samples will produce high-quality data is likely to be task-dependent (Krantz & Dalal, 2000). For tests developed and validated in the lab, Web versions have in some cases been shown to measure different constructs (see, e.g., Buchanan, Johnson, & Goldberg, 2005). Some tests developed in the lab might function best when specifically adapted for Web administration (Buchanan et al., 2005).

As in the lab, adequate quality assurance measures contribute to data quality (Reips, 2000, 2007). Here, we excluded Web participants who self-reported cheating, technical difficulties, inappropriate equipment, repeat performance, or very young/old age. Appropriate exclusion criteria are dependent on the research question and the type of experiment, however, and these particular exclusions may not be appropriate or sufficient for many types of experiments. Other researchers have emphasized additional variables that may be important for Web experiments, such as differences in computer environments, traffic sources, and technical expertise (Buchanan, 2007; Reips, 2002, 2007). Inclusion of participants using tablet computers or smartphones, for example, may impact data quality and should be considered when extending experiments from the lab to the Web. As with lab experiments, Web researchers must carefully consider potential confounding variables that could reduce data quality for any given test.

The present study does have important limitations. As the source of Web traffic may impact sample characteristics, data quality may vary depending on where the participants come from and how they are recruited. We did not collect information about how any particular participant came to our website (as recommended by Reips, 2002), so we were unable to see whether site of entry had any relationship with test scores or data quality. Personalized feedback may also have an effect on conscientiousness and types of participants when compared to other types of incentives. Finally, it is not known how variations in technical knowledge, computer type, and local environment may have affected the present results, as these characteristics were not measured (Buchanan, 2007; Buchanan & Smith, 1999; Reips, 2007). Given these potential differences between our tests and those of other research studies, our results should not be taken as evidence that data from self-selected Web samples are *always* of comparable quality to lab samples.

Rather, it is the responsibility of individual researchers to conduct their own assessments of data quality (Buchanan, 2007; Buchanan & Smith, 1999).

Our results indicate that, given basic quality assurance measures, self-selected Web samples are a potential source of high-quality data for cognitive and perceptual experiments, and can provide data similar to those collected in the lab or with more traditional methods of recruitment. As compared with lab-based experiments, Web-based experiments can yield very large and demographically diverse samples, enabling access to rare populations or low-base-rate characteristics, data-driven model building, speedy replication, and broader generalizability. Our results support efforts to take advantage of the technological and human resources available to behavioral scientists through Web-based experiments.

References

Australian Twin Registry. (n.d.). Australian Twin Registry, publications. Retrieved April 20, 2011, from www.twins.org.au/index.php?page=52

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42,* 241–251.

Birnbaum, M. H. (2000). Decision making in the lab and on the Web. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 3–34). San Diego, CA: Academic Press.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., et al. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant-stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology, 26,* 423–455.

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.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., et al. (2005a). Nonequivalence of online and paper-and-pencil psychological tests: The case of the Prospective Memory Questionnaire. *Behavior Research Methods, 37,* 148–154.

Buchanan, T., Johnson, J. A., & Goldberg, L. (2005b). Implementing a Five-Factor personality inventory for use on the Internet. *European Journal of Psychological Assessment, 21,* 115–127.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90,* 125–144.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2010). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

College Board. (2012). SAT percentile ranks for males, females and total group. Retrieved January 20, 2012, from http://professionals.collegeboard.com/profdownload/sat_percentile_ranks_composite_cr_m_w.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology, 24,* 419–430.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44,* 576–585.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15,* 373–378.

Germine, L., Cashdollar, N., Düzel, E., & Duchaine, B. (2011a). A new selective developmental deficit: Impaired object recognition with normal face recognition. *Cortex, 47,* 598–607.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011b). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118,* 201–210.

Germine, L. T., & Hooker, C. I. (2011). Face emotion recognition is related to individual differences in psychosis-proneness. *Psychological Medicine, 41,* 937–948.

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59,* 93–104.

Haworth, C., Harlaar, N., Kovas, Y., Davis, O., Oliver, B., Hayiou-Thomas, M., et al. (2007). Internet cognitive testing of large samples needed in genetic research. *Twin Research and Human Genetics, 10,* 554–563.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *The Behavioral and Brain Sciences, 33,* 111–135.

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods, 40,* 840–857. doi:10.3758/BRM.40.3.840

Kendler, K., Myers, J., Potter, J., & Opalesky, J. (2009). A Web-based study of personality, psychopathology and substance use in twin, other relative and relationship pairs. *Twin Research and Human Genetics, 12,* 137–141.

Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers, 29,* 264–269.

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). San Diego, CA: Academic Press.

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,* 105–117.

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of Web-delivered experiments: Can you trust the data? *Psychological Science, 11,* 502–506.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration Web site. *Group Dynamics: Theory, Research, and Practice, 6,* 101–115.

Pew Research Center. (2012a). Pew Internet, Internet adoption 1995–2011. Retrieved January 20, 2012, from www.pewinternet.org/Trend-Data/Internet-Adoption.aspx

Pew Research Center. (2012b). Pew Internet, Online activities, 2000–2009. Retrieved January 20, 2012, from www.pewinternet.org/Trend-Data/Online-Activities-20002009.aspx

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York, NY: W. H. Freeman.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). San Diego, CA: Academic Press.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49,* 243–256.

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. N. Joinson, K. Y. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). New York, NY: Oxford University Press.

Ronalds, G. A., DeStavola, B. L., & Leon, D. A. (2005). The cognitive cost of being a twin: Evidence from comparisons within families in the Aberdeen children of the 1950s cohort study. *British Medical Journal, 331,* 1306.

Senior, C., Phillips, M. L., Barnes, J., & David, A. S. (1999). An investigation into the perception of dominance from schematic faces: A study using the World-Wide Web. *Behavior Research Methods, 31,* 341–346.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100,* 330–348.

Uttl, B., Graf, P., & Richter, L. K. (2002). Verbal paired associates tests: Limits on validity and reliability. *Archives of Clinical Neuropsychology, 17,* 569–583.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Pearson.

Wellesley College. (n.d.). Wellesley College—Admission & financial aid—Testing. Retrieved January 20, 2012, from http://web.wellesley.edu/web/Admission/Apply/FAQ/testing.psml

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences, 107,* 5238–5241.