

Collecting response times using Amazon Mechanical Turk and Adobe Flash

Travis Simcox · Julie A. Fiez

Published online: 14 May 2013
© Psychonomic Society, Inc. 2013

Abstract Crowdsourcing systems like Amazon’s Mechanical Turk (AMT) allow data to be collected from a large sample of people in a short amount of time. This use has garnered considerable interest from behavioral scientists. So far, most experiments conducted on AMT have focused on survey-type instruments because of difficulties inherent in running many experimental paradigms over the Internet. This study investigated the viability of presenting stimuli and collecting response times using Adobe Flash to run ActionScript 3 code in conjunction with AMT. First, the timing properties of Adobe Flash were investigated using a phototransistor and two desktop computers running under several conditions mimicking those that may be present in research using AMT. This experiment revealed some strengths and weaknesses of the timing capabilities of this method. Next, a flanker task and a lexical decision task implemented in Adobe Flash were administered to participants recruited with AMT. The expected effects in these tasks were replicated. Power analyses were conducted to describe the number of participants needed to replicate these effects. A questionnaire was used to investigate previously undescribed computer use habits of 100 participants on AMT. We conclude that a Flash program in conjunction with AMT can be successfully used for running many experimental paradigms that rely on response times, although experimenters must understand the limitations of the method.

Keywords Response times · Crowdsourcing · Amazon Mechanical Turk · Adobe flash · ActionScript · Stimulus presentation · Web experiment · Rich media · Timing

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0345-y) contains supplementary material, which is available to authorized users.

T. Simcox (✉) · J. A. Fiez
Department of Psychology, University of Pittsburgh, Pittsburgh,
PA 15260, USA
e-mail: trs60@pitt.edu

Recently, there has been significant interest in conducting behavioral science research on Amazon’s Mechanical Turk (AMT; www.mturk.com). AMT is an online system in which *requesters* can open accounts and post a wide variety of tasks online, called *human intelligence tasks*, or *HITs*. *Workers* then browse the available HITs, and requesters can pay workers to complete the HITs.

AMT was not created with behavioral science in mind, but it is in many ways suitable for such applications. Mason and Suri (2012) provided a general guide for using AMT in behavioral research and emphasized some benefits of online experimentation using AMT. First, AMT provides a stable pool of participants, in contrast to undergraduate participant pools in which the supply of participants varies over the course of the semester or may be very limited at smaller universities. Second, AMT workers are diverse, as compared with undergraduate research pools. They come from a broad range of ages, ethnicities, economic backgrounds, and countries of origin. This allows for investigations of groups not present in undergraduate research pools or convenience samples, as well as generalizations of psychological phenomena to populations very different from undergraduate participant pools and to environments other than the lab. Heavy reliance on undergraduate research pools has drawn criticism that much of behavioral science is actually based on an unusual minority of the world’s population (e.g., Henrich, Heine, & Norenzayan, 2010). It is possible for individual requesters to send messages through AMT to workers that have previously worked for them and optionally include monetary bonuses with the messages. This allows for the use of performance incentives. Using these features, an experimenter could also develop multiple-component studies spanning multiple days or even deploy a demographic survey and then use the results to recruit a stratified or intentionally biased sample of AMT workers to represent a particular population. An additional benefit of AMT is the low cost of participant recruitment,

since many AMT workers participate for reasons other than monetary compensation. Finally, data collection can be very rapid. Buhrmester, Kwang, and Gosling (2011) received between 5.6 and 40.5 responses per hour to a set of surveys posted on AMT.

Most behavioral research currently listed on AMT uses survey-type instruments. This is not without cause: In a survey design, instruments can be written in HTML and posted directly on AMT, or they can be hosted using a ready-made commercial survey package like Qualtrics (e.g., Goodman, Cryder, & Cheema, *in press*; Paolacci, Chandler, & Ipeirotis, 2010), allowing a paper-and-pencil study to be converted to an experiment on AMT in a matter of hours. Many research questions do not lend themselves to a survey format, however. Of specific interest in this article are those paradigms that involve the collection of response times (RTs) from participants pressing keys in response to visual stimuli. Such paradigms have been instrumental in investigating a wide range of phenomena in the domains of perception, memory, attention, language, reasoning, problem solving, decision making, and movement control (Meyer, Osman, Irwin, & Yantis, 1988). Hewson, Laurent, and Vogel (1996) envisioned that the Internet would allow for collecting RTs from large and diverse participant pools in online experiments. A few such experiments have successfully collected data from many thousands of participants (e.g., Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Owen et al., 2010). These experiments did not use a system like AMT but, rather, used elaborate dedicated Web sites operating over a period of months. There is currently no easy route when it comes to collecting RTs online, since HTML provides no mechanism with which to do this. However, with the right approach, it is possible for experiments that gather RTs to enjoy the same benefits from AMT as survey research. The aim of the set of experiments presented here is to establish a method that experimenters can use with AMT to administer experimental stimuli and accurately collect RTs, despite the possibly large variation in the software, hardware, and computer use habits of AMT users.

A technology used for capturing RTs from AMT experiments must run inside of a Web browser, rather than being an external program downloaded by the worker. The AMT policies (AMT FAQ, 2012) prohibit requesters from requiring workers to download stand-alone software in order to complete a HIT. This precludes the use of standard stimulus presentation software that an experimenter would use in a lab. However, the policy allows requesters to use plug-ins and other rich media technology that run inside a Web browser. It is likely beneficial for experimenters to choose a rich media technology that is already ubiquitous. AMT experimental samples have some benefits over other types of samples, but they also score relatively high on measures of Internet and computer experience, understanding, and knowledge (Behrend, Sharek, Meade, & Wiebe, 2011). A

technology that, for example, requires a particular operating system, is used only by more advanced users, or is used only in offices could both reduce the size of the pool of possible participants and further bias it toward tech-savvy participants. Adobe's Flash fits these criteria of running inside the Web browser and ubiquity: Adobe claims that Flash Player is installed on over 98 % of Internet-connected computers (Adobe Developer Connection, 2012).

A suitable technology must have accurate timing, but it is not clear how accurate the timing must be. Flash Player runs in a virtual machine, which means programs running in Flash Player are isolated from system components like device drivers. This allows for information security and consistency across Web browsers and operating systems, but it also means that timing measurements must go through another layer separating an experimenter's program from the hardware that runs it. Several articles have described timing properties of Flash (e.g., Houben & Wiers, 2008; Reimers & Stewart, 2007; Schmidt, 2001), but the experiments were conducted using Flash Players based on an older virtual machine (AVM1). Newer versions of Flash Player support the new language ActionScript 3, and code written in ActionScript 3 runs on AVM2, a virtual machine that runs up to 10 times faster than AVM1 (Grossman & Huang, 2009).

After the release of AVM2-based software, Neath, Earle, Hallett, and Surprenant (2011) revisited the topic of chronometric experimentation in Flash. They found that RTs collected in a program implemented using Flash were still less accurate than those collected by MATLAB and Octave using Psychtoolbox. Additionally, no software can yield perfect timing accuracy, because the hardware does not allow for it. Plant and Turner (2009) investigated the timing abilities of commodity computer hardware. They found delays of up to 54 ms in monitors, 34 ms in keyboards, 49 ms in mice, and even 37 ms in speaker systems. Not only did the magnitude of timing errors vary by model of device, but their distributions also varied by device, precluding a solution of simply subtracting a constant value from measurements. Furthermore, neither equipment cost nor any other obvious factor translated to better timing abilities. Fortunately, perfect timing is not necessary. Brand and Bradley (2012) used a simulation to argue that even in a nonrobust experimental design, the amount of increased error due to technical variance that could be expected for a Web experiment over that of a lab-based experiment would not reduce statistical power substantially and that large sample sizes available on the Internet could provide an overall increase in statistical power.

The aim of Experiment 1 in this study was to describe the timing accuracy of an experiment developed in Adobe Flash under various conditions of computer use. This experiment was important in order to establish that the software and method are capable of accurate RT measurement before one

can expect to use it to replicate prior RT work. Specifically, it was hypothesized that timing accuracy in Adobe Flash meets the theoretical requirements necessary to detect a difference in RTs between experimental conditions with a small effect size. The aim of Experiment 2 in this study was to empirically determine whether the flanker effect, a classic effect in cognitive research, can be found using a Flash-based experiment posted on AMT and, if so, what considerations must be given to the sample size. Even if Experiment 1 revealed that the program implemented in Adobe Flash has accurate timing measurements, it would be conceivable that some aspect of the AMT system or AMT workers could make replication of laboratory findings difficult. Experiment 2 addressed this concern. It was hypothesized that the results of Experiment 2 would replicate a three-way interaction described by Nieuwenhuis et al. (2006, Experiment 1). Experiment 3 replicated a lexical decision task, in order to generalize the success or failure of the method described in this article beyond a single task. Specifically, it was hypothesized that the results of Experiment 3 would replicate a two-way interaction described by Lupker, Pera, and Davis (2008, Experiment 1b). The particular tasks used in Experiments 2 and 3 were chosen because both tasks had been well established in the cognitive science literature, both had manipulations that resulted in large classic effects on RTs that were expected to be easy to replicate, and the specific implementations of both tasks had effects on RTs that were smaller than the classic effects. In addition to the timing experiment and the two replications using AMT workers, some previously unexplored characteristics of AMT workers are described. When designing experiments to run on AMT, experimenters may want to take into consideration factors like the types of devices their participants are using or the environments their participants are currently in. These concerns were addressed through a questionnaire administered to the participants in Experiments 2 and 3.

Experiment 1

Experiment 1 investigated the timing accuracy of an Adobe Flash-based speeded decision task. It used an external timing apparatus for comparison of timing measurements, and it did not use human participants. The focus of the experiment was on the accuracy of stimulus presentation times, which is the major source of variability introduced by Flash into the RT measurements. The accuracy of keypress detection is addressed in the discussion.

Method

Experiment 1 was programmed in ActionScript 3 and compiled in Adobe Flash Professional CS5.5. The frame rate for

the Flash file was set to 60.00 FPS. On each trial, a white dot was presented inside a text field against a black background. When a keypress was detected by an event listener, the system time was captured by calling the Date() class in ActionScript, the contents of the text field were changed to no characters, the system time was again captured, and setTimeout() was used to call a function with a 1,000-ms delay. The function that was called by setTimeout() replaced the contents of the text field with the next item from an array, which in this experiment was always a dot. The system time was again captured. The difference between these two captured times—nominally, 1,000 ms—served as an interstimulus interval (ISI). Another function was called with a 100-ms delay. This second update did not visually change the stimulus, because it remained a dot. This function was included so that the ActionScript code used in this experiment would match the ActionScript code used for Experiment 2.

A phototransistor was placed on the monitor over the location of the dot. In this manner, the phototransistor detected the onset and offset of the stimuli while the experiment ran. It was monitored via the line-in jack on a laptop computer, which was continuously sampled at 8000 Hz. This allowed the timing information recorded by the Flash program to be compared with the intended presentation parameters and also to the high-resolution timing measured by the phototransistor.

In addition to the recording apparatus, two computers were used for testing. Because of the wide varieties of operating systems, processors, and Web browsers and the fact that new versions of each are released several times per year, it is not possible to exhaustively test the effect of each variety of each system component. Rather, two systems were used that could reasonably represent typical configurations. In 2012 and late 2011, most AMT workers used either Chrome or Firefox to access AMT (Simcox, 2012). Nearly half of AMT workers used Windows 7, although older versions of Windows also remained popular at the time of that data collection. We assumed that many users of outdated Windows versions have already upgraded or will upgrade soon. Because Linux was a rarity among AMT workers, we opted to include OSX as a second operating system choice. All OSX systems and most Windows systems ship with Intel processors, so both test systems used processors from the popular Intel Core series.

System 1 was a Dell Vostro running Windows 7 Enterprise Service Pack 1 on a 3.30-GHz Intel Core i3 processor with 4 GB of RAM. The display used with system 1 was a Samsung SyncMaster 740 N LCD display with a 60-Hz frame rate, common for computer displays and the default rate for Windows 7. The Web browser used for testing system 1 was Chrome 23.0.1271.97 m with Flash Player version 11.5.31.5. The stand-alone Flash player was

version 11.5.502.110. The software used to generate resource load on the CPU and RAM was Prime95 v27.7 build 2 (Woltman, 2012). System 2 was an Apple iMac model 11.3 running OSX 10.6.8 on a 2.8-GHz Intel Core i5 processor with 16 GB of RAM. The display in this system is an integrated LCD panel running at a 60-Hz frame rate. The Web browser used for testing system 2 was Firefox 14.0.1 with Adobe Flash Player 11.1.102.55. The stand-alone Flash player was version 11.3.300.268. Prime95 v27.7 build 2 was used to generate load. Both systems were tested under four conditions of resource load: low, medium, high, and maximum. In the low-load condition, the Flash program was the only window running in Firefox, with no other application software running. In the medium-load condition, the Flash program executed while Prime95 ran two processing threads of the torture test, an intense computational task. In the high-load condition, the number of threads running a torture test was increased to four. The medium- and high-load conditions were intended to mimic possible loads that occur if participants ran other software on their computers while participating in the experiment. In the maximum-load condition, Prime95 was instructed to use all system RAM to run four threads of the torture test, the process priority of Prime95 was manually raised, and the process priority of Flash Player was manually lowered. This was intended to force errors in timing to occur. The maximum-load condition was run to determine whether timing measurements collected by ActionScript could detect errors in presentation times controlled by ActionScript. If it were possible to detect errors in timing, it would allow experimenters to determine whether any participants are likely to have experienced inaccurate timing during the experiment by screening their data sets for participants with unusually high variability in the ISIs.

In a no-browser condition, the Flash program was executed in a stand-alone Flash Player rather than within a Web browser. Flash Player runs in a sandbox, so the browser and operating system both exist between Flash Player and the hardware. Versions of Flash Player used in browsers must conform to the browser's specifications, such as Chrome's Pepper Plugin API. Numerous anecdotes on discussion forums suggest that timing issues can arise when running a Flash program inside a Web browser that were not seen when running it in the stand-alone player. The no-browser condition was run to determine whether such a difference could be detected.

One hundred trials were run in each condition, and the first author served as the operator. It is critical to note that the tests here are intended to determine whether the magnitudes of timing errors make replications of chronometry experiments plausible. They are not intended to determine the effects of all possible variables. There is a virtually unlimited array of possible hardware and software configurations that could be

used to run Flash and, hence, could be used by AMT workers. In addition, the types of loads that could be running on an AMT worker's system are many, and the combination of ActionScript functions used here is one of many possible approaches to present stimuli, measure time, and record responses in ActionScript.

Results and discussion

The ISIs were all intended to be 1,000 ms. For the ISIs on each trial in each condition, the time elapsed as measured by the phototransistor output and the time elapsed as measured by the ActionScript output were recorded. Deviation from 1,000 ms in the phototransistor measurements represents error in presentation duration. Differences between the ActionScript measurements and the phototransistor measurements indicate error in the Flash application's ability to accurately record stimulus presentation duration. Histograms of the measurements of the ISIs from ActionScript and from the phototransistor are presented in Fig. 1. The means, standard deviations, minimum values, and maximum values of both measures, along with two-sided paired *t*-tests and 95 % CIs of the differences, are listed in Table 1. Although the focus of the comparisons is between the ActionScript measurements and the phototransistor measurements using two representative computer systems, a statistical comparison of these data between system 1 and system 2 is provided as [supplemental material](#).

As can be seen in the histograms in Fig. 1, times measured by the phototransistor fall into bins that are multiples of the 16.7-ms refresh rate. This is what should be expected if timing is imperfect, because when a key is pressed, the display cannot change instantly. Rather, the stimulus must remain at least until the display is updated at the next screen refresh. The low-load, medium-load, and high-load conditions were intended to represent typical computer use. In these three conditions collapsed across the two systems, a combined total of 80.5 % of intervals fell on the correct 1,000-ms screen refresh. Another 19.2 % of the intervals in these conditions deviated from 1,000 ms by a single screen refresh, and the remaining 0.3 % deviated by two refreshes. When the interval was less than 1,000 ms, the stimulus prior to the interval disappeared later than intended after the keypress, and when the interval was greater than 1,000 ms, the next stimulus appeared a refresh later than intended. It is also possible for both delays to happen on any individual trial, resulting in a 1,000-ms interval. It is unlikely that the variance in either process is multiple refreshes in these conditions, though, because that would result in a wider distribution.

It was not until load was abnormally high, in the maximum-load condition, that many outliers occurred and a wider distribution resulted. Levene's test was used to compare the variances in the low-load and the maximum-

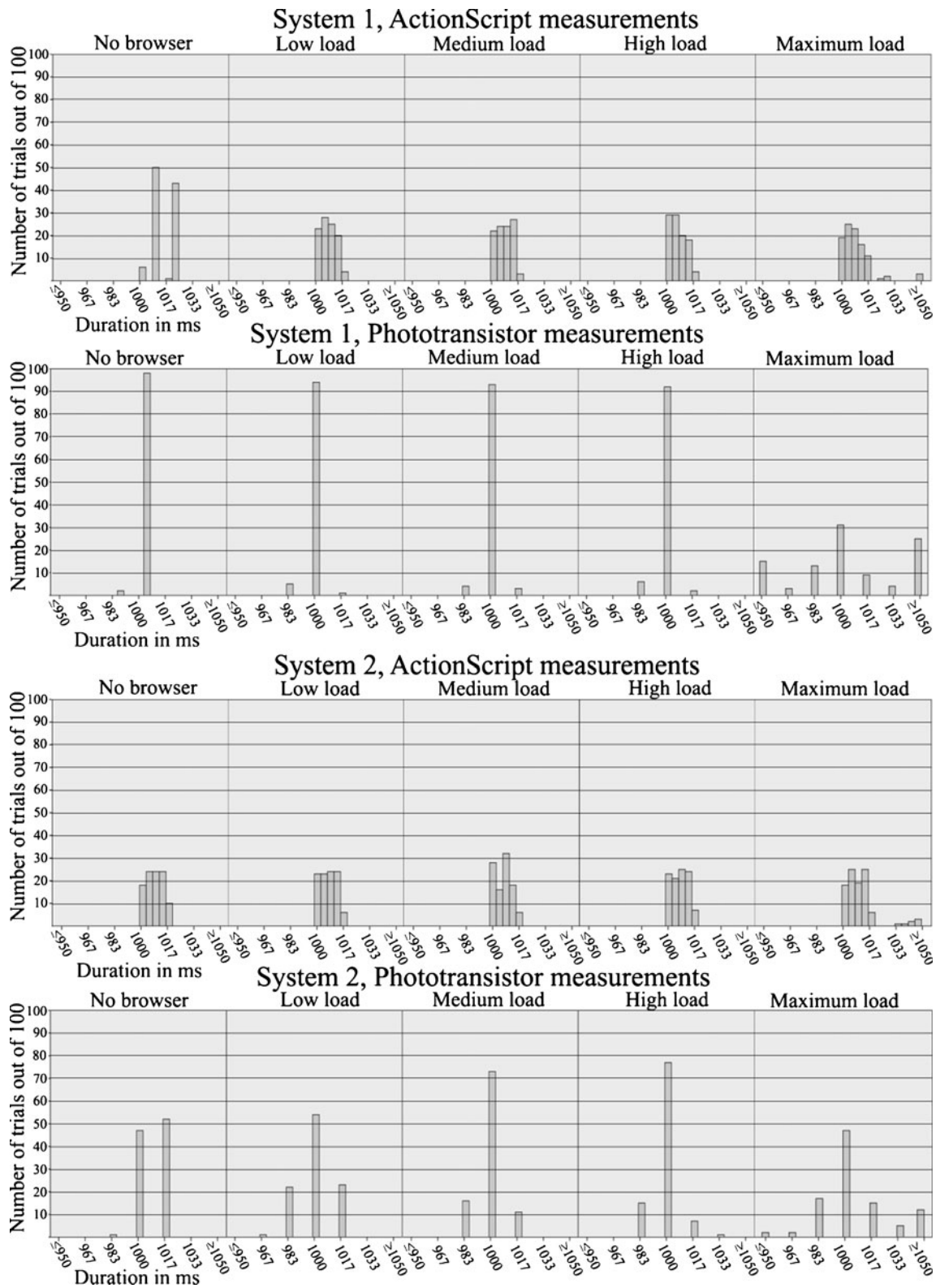


Fig. 1 Histograms of durations of interstimulus intervals for each condition, as measured by the phototransistor and by the ActionScript code. The x-axis is the measured duration in milliseconds

load conditions. The maximum-load condition had greater variance than the low-load condition in both the

phototransistor measurements, $F(1, 198) = 7.82, p = .006$, and the ActionScript measurements, $F(1, 198) = 7.93, p =$

Table 1 Durations of interstimulus intervals (in milliseconds) for each system in each condition

System	Condition	Actionscript values				Phototransistor values				Paired differences				<i>t</i> (99)	<i>p</i>	
		Mean	<i>SD</i>	Min	Max	Mean	<i>SD</i>	Min	Max	Mean	<i>SD</i>	<i>SEM</i>	95 % CI			UL
System 1	Stand-alone	1,014	5.98	1,000	1,020	1,003	2.35	987	1,004	11	5.42	3.50	9.29	11.76	16.93	<.001
System 1	Low load	1,008	4.59	1,000	1,016	1,001	4.04	985	1,019	7	4.93	0.12	5.51	7.76	11.74	<.001
System 1	Medium load	1,008	4.74	999	1,016	1,002	4.43	985	1,019	7	4.42	0.12	5.13	7.26	11.54	<.001
System 1	High load	1,007	4.78	999	1,016	1,001	4.68	985	1,018	6	5.06	0.12	4.82	7.00	10.75	<.001
System 1	Max load	1,017	81.31	999	1,815	1,027	183.38	69	1,935	91	200.06	0.75	-53.46	33.77	-0.45	.655
System 2	Stand-alone	1,009	8.71	984	1,018	1,009	4.93	1000	1,017	5	3.99	0.00	-1.12	1.56	0.32	.747
System 2	Low load	1,008	4.86	999	1,018	1,000	11.75	967	1,018	11	8.66	1.00	5.28	10.08	6.35	<.001
System 2	Medium load	1,008	4.99	999	1,016	1,000	8.66	984	1,018	10	7.86	0.00	5.63	9.59	7.63	<.001
System 2	High load	1,008	5.09	999	1,017	1,000	8.50	984	1,034	10	7.54	0.00	6.43	9.97	9.20	<.001
System 2	Max load	1,012	13.13	999	1,092	1,191	1,255.10	784	13,144	207	1251.00	0.00	-428.53	69.72	-1.43	.156

.005. This suggests that if a substantial number of inaccurate trials are present for a participant, the problem can be detected by using ActionScript to measure the duration of the ISI for each trial. Experimenters could determine cutoffs at which they could exclude participants whose equipment generated too many inaccurate intervals.

For both systems, the longer ISIs measured by ActionScript, when compared with the phototransistor (as shown in Table 1), were offset by shorter stimulus display durations measured by ActionScript than measured by the phototransistor. Additionally, the recorded stimulus presentation time always occurred immediately after detection of the keypress, despite the fact that actual presentation took place in intervals of screen refreshes. There are a couple of factors that contributed to these timing shifts. One factor was detected postexperiment through a combination of observing the output of different variations of the ActionScript code and profiling Flash Player during execution using the aid of the advanced telemetry features of Adobe Scout. Much of the variation in the ActionScript intervals was narrowed down to the setTimeout() method taking slightly longer than the specified interval to run. As a result, some code modifications were explored. In one modification, a 990-ms interval was requested from setTimeout(), followed by a loop that constantly queried the system time until 1,000 ms had passed and then executed the code to update the display. This method resulted in an average ActionScript-recorded interval of $M = 1,000.2$ ms ($SD = 0.84$ ms) and in higher CPU utilization (3 % vs. 1 % on system 2) immediately before stimulus presentation than did setTimeout() alone. The second factor involved in the timing shift is that Flash operates in a security sandbox at a nominal frame rate that is set prior to compiling the project, and the resulting program does not have the ability to synchronize its refresh schedule with the actual refresh time of the user’s monitor. No matter how accurately ActionScript requests a screen update, it does not know the exact time that the display actually changes.

Stimulus presentation times have so far been the focus of this experiment, but RT collection also depends on the accuracy of a second component: the detection of the keypress. Some researchers have used highly accurate electronic devices to press keyboard keys at known intervals and compared the known time with the time that the computer recorded the keypress (e.g., Neath et al., 2011; Plant & Turner, 2009). This method reveals the keypress timing accuracy of the combination of the hardware and software. Because keypress timing error is already an established phenomenon in the literature and is not unique to Web-based experiments, we sought to determine the accuracy of Adobe Flash in collecting keypress times when keyboard hardware was removed as a factor. Accuracy of keypress detection was assessed using a macro implemented in Logitech Profiler 5.10 (Logitech International S.A., Romanel-sur-Morges, Switzerland). The macro recorded the operator’s responses to one run of the

100-trial experimental task. The macro was played back 25 times on system 1 while E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA) recorded keypress times and the experiment was idle in Chrome in the background. The recorded timings of the keypresses in each of the runs were very similar, with an average interitem correlation of .999949 and a pooled standard deviation of 0.76 ms, demonstrating that the macro was capable of reproducing a series of responses with very high reliability. The 25 runs in E-Prime were averaged in order to serve as a “known” value of the time of each keypress in the series. The macro was then run in each condition. The absolute value of the difference between the known and measured time for each keypress was calculated. The average error for trials was 0.26 ms in the low-load condition, 0.37 ms in the medium-load condition, and 0.31 ms in the high-load condition. The average error for the no-browser condition was 3.22 ms by the measure described above, and the macro terminated over 200 ms more slowly in the no-browser condition than in the other conditions. This discrepancy was small yet reproducible, but its cause is unknown. Because AMT workers cannot be required to run programs outside of their Web browser, it is not an issue that would arise in the course of an AMT Web experiment. In the maximum-load condition, inferences about the accuracy of keypress timing were not made because the stimulus presentation could not keep pace with the macro.

Because neither Logitech Profiler nor E-Prime runs on system 2, an identical macro was compiled in the C programming language and was executed on system 2 while Firefox ran the experimental task. Twenty-five runs were recorded to create a known value, analogous to the known value calculated for system 1. The measurements in the 25 runs had an average interitem correlation of .999981 and a pooled standard deviation of 0.46 ms, demonstrating that the macro was capable of reproducing a series of responses with very high reliability on system 2 as well. The 25 runs were averaged in order to serve as a “known” value of the time of each keypress in the series. As was expected, the known values collected under system 1 and the known values collected under system 2 were nearly identical and correlated with a Pearson's r exceeding .999999.

As with system 1, the macro was then run in each condition on system 2. The absolute value of the difference between the known and measured times for each keypress was calculated. The average error for trials was 0.39 ms in the low-load condition, 0.40 ms in the medium-load condition, and 0.64 ms in the high-load condition. The average error for the no-browser condition was 0.41 ms. In the maximum-load condition, inferences about the accuracy of keypress timing were not made, because the stimulus presentation could not keep pace with the macro.

The results of the keypress timing data suggest that the amount of RT measurement error attributable to Flash's

detection of keypresses reported by the operating system is negligible. Plant and Turner's (2009) finding that keyboards themselves can introduce error is important to consider, but there is no reason to expect that a Web experiment using Adobe Flash would have less accurate detection of keypresses than a lab-based experiment using standard stimulus presentation software and a keyboard as an input device.

The inability of ActionScript to detect the actual display time of the stimulus means that RTs measured in a Flash-based experiment using the approach here will not be quite as accurate as those measured using a software package that can synchronize presentation to the refresh of the monitor. However, this is unlikely to be a problem for most experimental paradigms running under most conditions. Measurement error of the magnitude seen here falls well within the bounds considered by Brand and Bradley (2012) in their investigation of statistical power using simulated Web experiments. In their simulations, they found that a random 10- to 100-ms delay on each trial resulted in only 1 % loss of statistical power when detecting a manipulation that results in a 20-ms difference in response times. This is because the error due to technical variance is small, as compared with variance in human RT distributions. The measurement error found here also falls within the bounds considered by Ulrich and Giray (1989). Ulrich and Giray demonstrated mathematically that, contrary to what people may intuitively expect, a clock that quantizes RTs in 32-ms intervals could detect the effect of a manipulation that results in a 20-ms change in RT about as well as a perfect clock could detect the same effect. Thus, the method used in this experiment should be able to reliably detect a difference in RTs between conditions that is at least as small as 20 ms. With only a 1 % loss in statistical power for a 20-ms difference, it seems likely that an even smaller difference could be detected with a reasonable amount of power. However, neither Ulrich and Giray nor Brand and Bradley investigated differences smaller than 20 ms. Brand and Bradley showed that with typical human RT distributions, 20 ms corresponded to what Cohen (1988) described as a small effect size.

Experiment 2

Here, we replicated the flanker task described by Nieuwenhuis et al. (2006, Experiment 1). In this task, the participant rapidly categorizes a series of target stimuli that appear in a known position in a visual display, while ignoring the rest of the display. These stimuli are arrows pointing left or right, and the participant responds by pressing a key with their corresponding left or right hand. On congruent trials, targets are surrounded by stimuli (flankers) that are identical to the target. On incongruent trials, the flankers point in the opposite

direction as the target. The incongruence causes conflict, which results in longer RTs and increased error rates. Incongruence on a particular trial can also influence a participant's response to the next trial in the sequence. When a sequence of trials contains two consecutive incongruent trials, the slowing effect of conflict is diminished on the second of the two incongruent trials. This phenomenon is called conflict adaptation, and it results in a two-way interaction effect in the data. Nieuwenhuis et al. found that the conflict adaptation effect was present only for trials that required the same response key as the previous trial, represented by a three-way interaction effect in the data.

Method

Participants

A listing (HIT) was posted on the AMT Web site looking for adult English speakers to participate in an experiment in exchange for \$0.70. The HIT specified that participants must be free of distractions and must have a recent version of Adobe Flash Player. Using the requester interface of AMT, qualifications were set for the HIT. These qualifications made it so the HIT was set to be visible and available only to those participants whose location was identified by Amazon as “United States,” who had completed at least 50 prior tasks on AMT, and whose prior work on AMT had been accepted by requesters at least 95 % of the time. Although we are not aware of any empirical investigations into the effects of these qualifications, AMT recommends the use of qualifications in order to reach high-quality workers (The Mechanical Turk Blog, 2012). The listing was visible until all of the posted HITs had been accepted (approximately 5 h). Upon clicking the link in the HIT, participants saw a screen describing the nature of the research and were taken to Experiment 2 if they consented to participate.

One hundred participants accepted the HIT and completed Experiment 2 (58 female, 38 male, 4 unspecified gender). Ages ranged from 18 to 70 years ($M = 38$, $SD = 14$). Seventeen participants reported being currently enrolled as university students. When asked for their highest level of education, 5 reported no degree, 25 had completed secondary school, 21 had some college, 32 had a Bachelor's degree, 12 had a Master's, and 5 had a Doctorate. Despite setting the HIT to be available only to AMT workers in the United States, 2 participants connected from non-US Internet service providers. Ninety-seven reported a native language of English, with 8 reporting fluency in a second language. All participants reported accessing the experiment from either a desktop computer ($n = 36$) or a laptop ($n = 64$) with a screen size larger than 8 in. Sixty-six reported having a keyboard with a number pad, although it was not used in

this study. Twenty-one reported being regular players of action video games. Ninety-two participants reported doing nothing else during the experiment, while 5 reported listening to music and 3 reported watching television. Ninety-four reported that they participated from their home. Forty-eight reported being in a silent environment, 43 reported “occasional noise,” 9 reported “frequent noise,” and none described their current environment as “very loud.” According to the user agent strings automatically supplied by the participants' browsers when they loaded the experiment, 18 systems were running Windows XP, 16 Windows Vista, 55 Windows 7, 1 Windows 8, and 10 OSX. Thirty-three were using the Chrome Web browser, 51 were using Firefox, and 16 were using Internet Explorer.

Procedure

The experimental session consisted of three components. All participants first completed a flanker task, which is described in this experiment. The flanker task was followed by one of two lexical decision tasks: a replication of Lupker et al.'s (2008) Experiment 1b, which is reported in Experiment 3, or an alternative lexical decision task of the same duration. The alternative lexical decision task was not a replication of existing work and is not presented here. The testing session then ended with a brief questionnaire. Participants were informed that their responses to the questionnaire would not affect payment, in order to avoid biasing results. According to the requester Web interface of AMT, the combined time of informed consent, both experimental tasks, the questionnaire, debriefing, and requesting payment took participants an average of 17 min. The experimental program loaded in its entirety by the time a participant consented and then sent data back to the server only after the questionnaire was completed. This ensured that the type and speed of a participant's Internet connection did not have an impact on a participant's experience or on data collection. Although preloading the entire experiment at the beginning and then transmitting the entire data set at the conclusion are simple details of the program design, they are important. An early pilot that communicated with the server between trials resulted in less consistent stimulus presentation on slower Internet connections.

The flanker protocol used in this study was designed to mimic the protocol in Nieuwenhuis et al. (2006, Experiment 1), but with a total of 8 rather than 17 blocks to keep the total experiment duration in line with the experimental tasks typically seen on AMT. On each trial of the flanker task, participants were presented with “<<<< <<<<” or “>>>> >>>>” in white text against a black background in the center of the window. After 100 ms, either “<” or “>” appeared in the center of the existing stimulus, resulting in four possible combinations of arrows, two of which were congruent combinations of stimuli and two of

which were incongruent. Each of the four possible combinations occurred 10 times per block, resulting in a total of 320 trials. Participants were instructed to press the “A” key to indicate that the center arrow pointed to the left or the “L” key to indicate that it pointed to the right on each trial. Before beginning the task, they were also shown two example stimuli—one congruent and one incongruent—along with the correct keypresses for these examples. Upon participants’ responses, a stimulus vanished, and the next stimulus appeared 1,000 ms later. Participants were instructed to respond as quickly and accurately as possible. Between blocks, participants were shown the number of blocks remaining and were prompted to press a button when they were ready to continue to the next block. The task was programmed in ActionScript 3 in a manner similar to that in Experiment 1. As in Experiment 1, ActionScript recorded the system time after each keypress and after each update of the text field.

Results and discussion

Before proceeding with the analysis of behavioral data, the ability of Turkers’ computers to keep accurate timing was gauged. To do this, the durations recorded by ActionScript for each of the 1,000-ms ISIs from all participants were calculated as in Experiment 1. Boxplots of the measurements from each participant are shown in Fig. 2, ranked from lowest to highest variance. These measurements are not dependent on the participants’ RTs but, rather, on the timing capability of their computers while they participate. A total of 41 measurements, or 0.13 %, exceeded 1,034 ms. The participant whose equipment showed the greatest variance had ISIs on 4.4 % of trials that exceeded 1,034 ms. These data suggest that the equipment used by most participants can operate with fairly consistent, although imperfect, timing. Using this strategy of timing fixed-duration intervals, experimenters could decide on criteria to exclude

participants whose equipment had timing reliability poorer than a chosen cutoff.

Overall error rates of each participant on the flanker task were calculated. Most participants ($n = 87$) were below 20 % overall error rate, but 13 participants clearly stood out with near-chance ($n = 12$) and above-chance ($n = 1$) error rates. These 13 participants were excluded from the ANOVA, but in an effort to understand the factors that may influence the performance of AMT participants, their data patterns were examined in more detail. One of these 13 participants had a 93 % error rate, which is exceedingly unlikely to occur by chance (chance level is 50 %). A reasonable explanation is that this participant had the keyboard keys backward when responding to the task and would have achieved a 7 % error rate if the instructions were to use the reversed keys. To attempt to explain the cause of the elevated error rates (43 %–51 %) in the other 12 participants, error rates were recalculated as if the goal was to respond to the flankers rather than the targets. If this had been the goal, 11 of these 12 participants would have achieved error rates under 20 %. The remaining participant also had unusually high error rates in another task, unusually short RTs, and no discernible pattern in responding. This participant was probably not putting effort into the task.

Each trial was classified by congruency (if the target arrow pointed the same direction as the flankers), the congruency of the preceding trial, and response type (if the target arrow pointed the same direction as the previous target arrow; repetition vs. change). For the RT data, incorrect responses, responses following incorrect responses, responses shorter than 200 ms, responses greater than 2 *SD* over the participant’s mean RT, and the first trial of each block were excluded from analysis, for a total of 11.63 % of data that were excluded from RT analysis. For the error rate analysis, only the first trial from each block was excluded from analysis, on the basis that it could not be categorized by previous trial type.

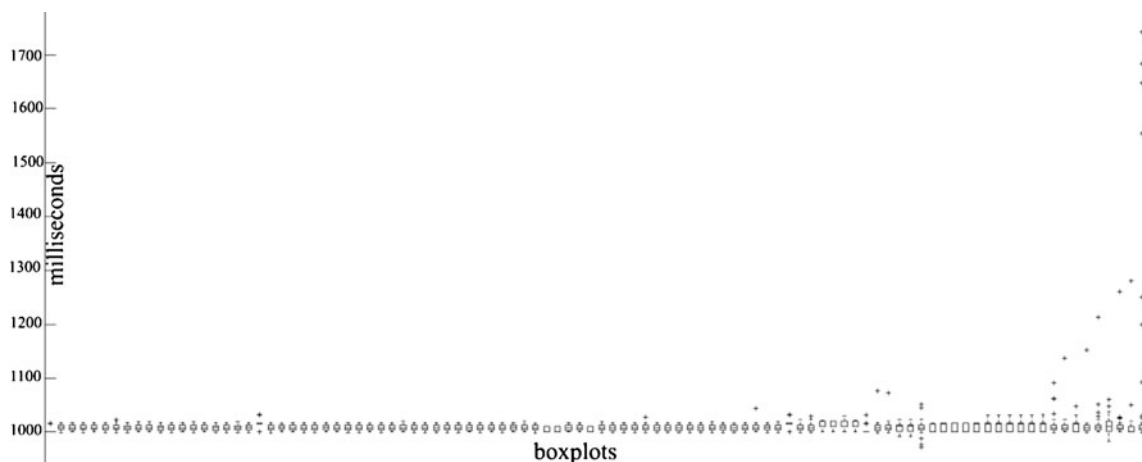


Fig. 2 Durations of intervals recorded by ActionScript for each Turker, ordered by variance

A pair of 2 (response type: repetition vs. change) \times 2 (current trial type: congruent vs. incongruent) \times 2 (previous trial type: congruent vs. incongruent) factorial repeated measures ANOVAs were conducted, one for RTs and one for error rates. The three factors were all within-subjects fixed effects, and participant was a random effect. The F -values of the ANOVAs of these data and those reported by Nieuwenhuis et al. (2006, Experiment 1) are listed in Table 2, and the similarity between Nieuwenhuis et al.'s results and the present results is clear in Fig. 3. Along with the F statistics are computed p -values, partial omega-squared effect size estimates, and 95 % confidence intervals of the partial omega-squared effect size estimates. Partial

omega squared was chosen as an appropriate measure of effect size to compare the data sets considered here because its value is not biased by the differences in sample size (e.g., Fowler, 1985). The effect size confidence intervals were calculated on the basis of Smithson's (2001) iterative method of finding the bounds of the noncentrality parameter for the noncentral F distribution.

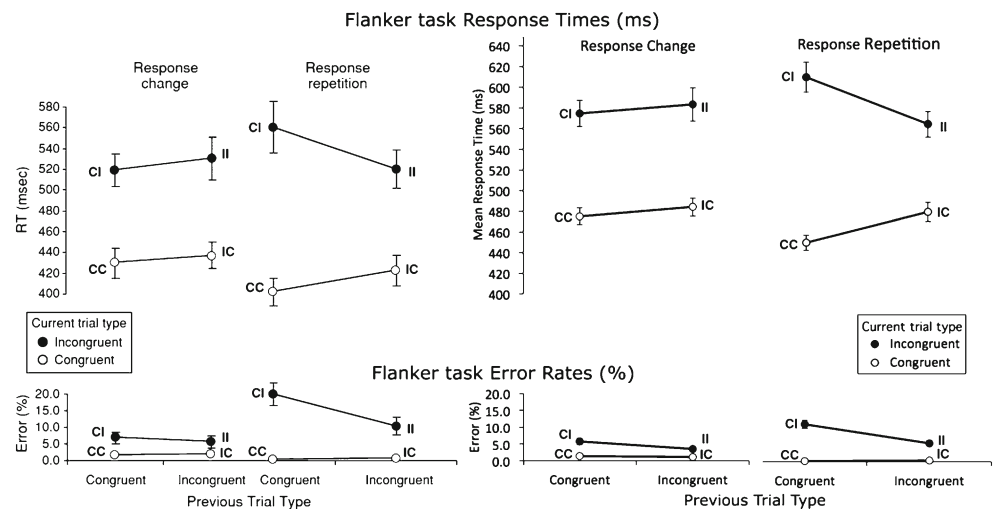
The ANOVAs showed main effects of current trial type (flanker effect) on RTs and on error rates in both studies. The two-way interaction effects of current trial type and previous trial type in the RT and error rate data were found, which is typically interpreted as evidence of conflict adaptation in this task. There was also a three-way interaction for

Table 2 Comparisons of effects found by Nieuwenhuis et al. (2006, Experiment 1) and effects found in present data

Factor	Response times									
	Nieuwenhuis					Present				
	$F(1, 13)$	p	ω_p^2	ω_p^2 95 % CI		$F(1, 86)$	p	ω_p^2	ω_p^2 95 % CI	
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Current trial type	153.1	<.001	0.910	0.777	0.956	163.42	<.001	0.649	0.536	0.734
Previous trial type	0	1.000	0.000	0.000	0.000	4.16	.045	0.035	0.000	0.156
Response type	0.1	.757	0.000	0.000	0.252	0.03	.856	0.000	0.000	0.038
Current \times previous	15.1	.002	0.485	0.120	0.739	29.30	<.001	0.243	0.110	0.394
Response \times previous	7.8	.015	0.312	0.016	0.639	20.15	<.001	0.179	0.062	0.330
Current \times response	11.1	.005	0.402	0.060	0.693	84.30	<.001	0.486	0.345	0.605
Current \times response \times previous	8.1	.014	0.321	0.020	0.645	52.21	<.001	0.368	0.221	0.506
Response repetition only										
Current \times previous	24.2	<.001	0.607	0.250	0.804	70.85	<.001	0.443	0.298	0.569
Response change only										
Current \times previous	0.1	.757	0.000	0.000	0.252	6.18	.015	0.056	0.002	0.187

Factor	Error rates									
	Nieuwenhuis					Present				
	$F(1, 13)$	p	ω_p^2	ω_p^2 95 % CI		$F(1, 86)$	p	ω_p^2	ω_p^2 95 % CI	
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Current trial type	21.9	<.001	0.582	0.219	0.791	62.63	<.001	0.412	0.266	0.544
Previous trial type	36.3	<.001	0.702	0.375	0.852	48.78	<.001	0.352	0.206	0.492
Response type	7	.020	0.286	0.007	0.623	15.22	<.001	0.139	0.038	0.288
Current \times previous	24	<.001	0.605	0.248	0.803	44.62	<.001	0.331	0.186	0.474
Response \times previous	7.8	.015	0.312	0.016	0.639	9.96	.002	0.092	0.014	0.235
Current \times response	11.1	.005	0.402	0.060	0.693	45.26	<.001	0.335	0.189	0.477
Current \times response \times previous	9.8	.008	0.370	0.042	0.674	25.21	<.001	0.216	0.089	0.367
Response repetition only										
Current \times previous	28.7	<.001	0.649	0.305	0.825	57.70	<.001	0.392	0.245	0.527
Response change only										
Current \times previous	0.8	.387	0.000	0.000	0.359	11.41	.001	0.106	0.020	0.250

Fig. 3 Left: Reprinted from Nieuwenhuis et al. (2006, Experiment 1). Mean response times (RTs) and error rates reported in Nieuwenhuis et al. (2006, Experiment 1). Right: Mean RTs and error rates for each trial category found in the AMT data. Error bars represent standard errors of the means



current trial type, response type, and previous trial type in both the RT and error rate data, because the conflict adaptation effect was significantly larger for response repetition trials than for response change trials. These findings parallel those of Nieuwenhuis et al. (2006, Experiment 1). In the RT data, the size of the flanker effect was larger in Nieuwenhuis et al.'s data than in the present data, although the size of the three-way interaction effect was similar in both data sets. Effect sizes in the two data sets may differ due to the difference in populations. AMT workers are more diverse in several ways than university undergraduates. They are also participating in more varied, less controlled conditions than lab participants. The effects of this are not necessarily obvious. For example, Paré and Cree (2009) found that AMT users rated images of objects as more familiar than did lab participants and suggested that the difference in familiarity ratings was because AMT users were participating in a more familiar environment than were lab participants.

A pair of 2 (current trial type: congruent vs. incongruent) \times 2 (previous trial type: congruent vs. incongruent) factorial repeated measures ANOVAs were conducted, one for RTs and one for error rates, on data from only the response repetition trials. Another pair of ANOVAs was conducted on data from only the response change trials. The response repetition trials showed a conflict adaptation effect, as evidenced by a two-way interaction, in both the RT and error rate data, as in Nieuwenhuis et al. (2006, Experiment 1). However, data from the response change trials revealed a two-way interaction in the RT and error rate data as well. These two-way interactions in the response change trials were not found in Nieuwenhuis et al. The interaction effect in the error rates for response change trials was smaller than the interaction effect in the response repetition trials, but it was in the direction consistent with conflict adaptation. The interaction effect in the RTs of the response change trials was smaller than the interaction effect in the RTs of the response repetition trials, and it was in the opposite direction of conflict adaptation. The emergence of

these significant interactions in the present data may be related to the fact that the participants in the present study were generally slower and more accurate than the participants in Nieuwenhuis et al., despite the similarities in the instructions given to participants and stimulus presentation parameters between the two experiments.

Planned contrasts described the magnitude of the overall conflict effect and the magnitude of the conflict adaptation effect in response repetition trials. The effect of conflict was significant: Congruent trials received responses 111 ms faster than incongruent trials (95 % CI [93 128]). On incongruent trials on which the correct response was a repetition of the previous response, the increase in RT due to the previous trial being incongruent was 34 ms (95 % CI [23 44]).

To complement the effect size measures, a resampling procedure was used to describe the required number of participants needed to detect the three-way interaction in the RTs. The effect size measures are useful for comparing the two data sets with the same design, but less useful in determining the sample sizes needed to detect effects. In a one-way ANOVA that meets a standard set of assumptions, estimating a required sample size from a measure of effect size is as simple as referring to a table. However, repeated measures from participants result in correlated measurements that violate the assumption of independent observations. If the correlation is known, free software like G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) can correct for it when estimating required sample size. However, this correlation is generally not included when reporting *F*-tests in published work. Furthermore, the present experiment is a 2 \times 2 \times 2 repeated measures design, resulting in multiple correlations, and there is no reason to suppose that, for example, the effect of congruency results in the same correlation at both levels of response type.

The resampling procedure shows at a glance the results obtained when small random samples of AMT users come forth from the large number that are online at any given

time, a situation assumed to occur in research using AMT. For a concise and accessible discussion of the advantages and criticisms of resampling methods, see Yu (2003). In order to conduct the power analysis, the ANOVA model described above was implemented in the R statistical computing environment (R Developmental Core Team, 2011). For each n from 2 to 50, n participants were randomly sampled with replacement from the 87 participants in the data set, and the model was fit to their RTs. This process was repeated 10,000 times, and the proportions of the models in which the three-way interaction was significant at alphas of .05, .01, and .001 are plotted in Fig. 4. The results suggest that a powerful analysis can be conducted on AMT using sample sizes that are comparable to those used in laboratory experiments.

Experiment 3

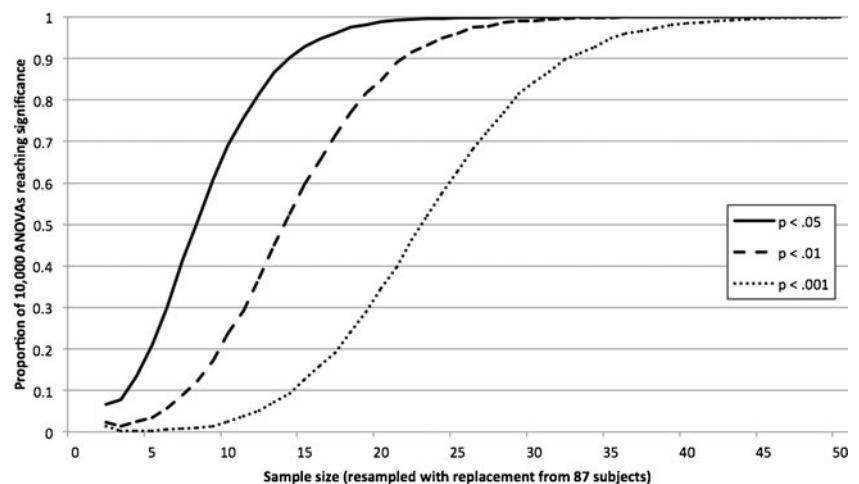
In a lexical decision task, the participant responds to a string of letters, indicating whether the strings form real words (Meyer & Roger, 1971). Lupker et al. (2008, Experiment 1b) used four methods of constructing nonwords. They found that nonwords constructed by transposing letters of a base word are responded to more slowly than nonwords constructed by replacing letters of a base word. This effect was larger for nonwords constructed by transposing consonants, as compared with those constructed by transposing vowels.

Method

Participants

The last 30 participants who participated in Experiment 2 also participated in Experiment 3 (19 female, 10 male, 1 unspecified gender). Ages ranged from 18 to 70 years ($M = 42$, $SD = 16$).

Fig. 4 Statistical power of three-way interaction as a function of sample size and significance criterion



Procedure

On each trial, a fixation point appeared for 500 ms, followed by an uppercase target that remained on the screen until response. Participants were instructed to press the “L” key if the target spelled a real English word or press the “A” key if it was a nonword and were instructed to respond as quickly and accurately as possible. The use of the left hand for nonword and right hand for word (not counterbalanced) followed the procedure of Lupker et al. (2008). Half of the 160 targets that each participant saw were words. Of the nonwords, 20 were generated by transposing two nonadjacent vowels, 20 by replacing two vowels with other vowels, 20 by transposing two nonadjacent consonants, and 20 by replacing two consonants with other consonants. Lists were counterbalanced across four groups of participants so that each participant saw all word stimuli and one nonword generated from each word. Further details and theoretical considerations of the stimuli are explained in Lupker et al.’s work; for this replication, it is important only that their manipulations resulted in different effect sizes. Each participant was randomly assigned to one of the four counterbalancing groups when they loaded the Flash file. This assignment method does not result in equal-sized groups. Because data are collected from several participants simultaneously and because clicking the link to the experiment does not guarantee that a participant will complete or even begin the experiment, there is no obvious way to ensure equal group sizes. The task was programmed in ActionScript 3 in a manner similar to Experiment 1 and Experiment 2, but with only one stimulus presented on each trial.

Results and discussion

The data trimming and ANOVAs followed that in Lupker et al. (2008, Experiment 1b), including the removal of the item “CEMERONY.” Incorrect responses (7.7 % of the total nonword trials and 2.7 % of the total word trials) and RTs

greater than 1,500 ms (10.7 % of the total nonword trials and 4.4 % of the total word trials) were excluded from the RT analysis. The 1,500-ms cutoff affected more nonwords in the present study than in Lupker et al. (12.7 % of nonword trials). For both the subjects analysis and the items analysis, a pair of 2 (alteration: transposition vs. replacement) \times 2 (letter type: consonants vs. vowels) \times 4 (counterbalancing list) repeated measures ANOVAs with participants as a random effect was conducted for RTs and for error rates for nonword stimuli only. Additionally, a power analysis was calculated. Words versus nonwords *t*-tests on RTs and error rates were calculated by collapsing across items to determine whether the classical lexical decision effect was present.

The initial word versus nonword comparison of RTs showed that responses to words were faster ($M = 821$, $SD = 109$, 95%CI[780 861]) than responses to nonwords ($M = 876$, $SD = 123$, 95%CI[830 922]), $t(29) = 4.63$, $p < .001$, $d = 0.85$. This is the classical RT effect found in lexical decision paradigms. Although the mean RT to words found by Lupker et al. (2008, Experiment 1b) of 799 ms is slightly shorter than the mean found here, it falls within the 95 % confidence interval. The comparison of error rates showed that error responses were less common for words ($M = 3.71$, $SD = 2.75$, 95%CI[1.68 3.73]) than for nonwords ($M = 7.67$, $SD = 6.32$, 95%CI[5.31 10.03]), $t(29) = 3.83$, $p < .001$, $d = 0.70$. In contrast, Lupker et al. found a higher error rate of 7.7 % for word trials.

F-tests of the ANOVAs on RTs and error rates are presented in Table 3, along with the corresponding *F*-values found by Lupker et al. (2008, Experiment 1b). Computed *p*-values, partial omega-squared effect size estimates, and 95 % confidence intervals for the partial omega-squared values are presented. The ANOVA on the RTs showed main effects of alteration and letter type and a significant interaction of these factors. Similarly, for error rates, there were main effects of alteration and letter type and a significant interaction of these factors. Table 4 shows values from the subject analysis ANOVA in milliseconds.

The difference in participant responses between the effect of replacing letters in a stimulus versus transposing letters in that stimulus to create nonwords is what Lupker et al. (2008) called the “transposed-letter effect.” The effect is replicated here in the error rate and RT data.

On the basis of the same reasoning as in Experiment 2, a resampling procedure was used to describe the required number of participants needed to detect the interaction effect of alteration and letter type in the RTs. The ANOVA model described above was implemented in the R statistical computing environment (R Developmental Core Team, 2011). For each n from 3 to 50, n participants were randomly sampled with replacement from the 30 participants in the data set, and the model was fit to their RTs. This process was repeated 10,000 times, and the proportions of the models in

which the two-way interaction was significant at alphas of .05, .01, and .001 are plotted in Fig. 5.

General discussion

The results of Experiment 1 showed that collecting RT measurements from AMT workers using Adobe Flash should be feasible, and the results from Experiment 2 and Experiment 3 provide evidence that the method is valid and useful. The data obtained from participants using programs built in Adobe Flash followed patterns that have previously been described in lab-based experiments, demonstrating that, at least for some paradigms, AMT is a valid source of RT data. Some others may not be appropriate for administration in the manner explored here or, at least, would require additional validation. Because the experimenter does not have access to information about screen refreshes and presentation time is not accurate to the individual screen refresh, a paradigm that relies on priming using extremely brief stimulus durations (i.e., one or two refreshes, or 33 ms) would fall into this category. Because fewer than half of participants reported being in a silent environment, experimenters should be careful when using auditory stimuli. Mayo, Aubanel, and Cooke (2012) instructed participants recruited using AMT to listen to words using headphones in a quiet environment but found significantly lower accuracy than in laboratory participants. Potentially, headphone requests may not eliminate the detrimental effects of a noisy background experienced by some AMT workers. As another practical point, in the data here, participants favored laptops over desktop computers. This would pose a problem for an experimental paradigm that relies on collecting RTs using the mouse as an input device.

In addition to technological considerations, the characteristics of the population are important. The participants in the present experiments were older than the participants in the lab-based studies that these experiments were modeled after and had a variety of educational backgrounds. These sorts of differences have been described in past work on AMT (e.g., Behrend et al., 2011; Berinsky, Huber, & Lenz, 2012), and here they likely had an effect on RTs and error rates. At one point in AMT’s history, all workers were American, but some surveys have found that most workers now reside outside the United States (Ipeirotis, 2010), introducing more variability in language background and other factors. While AMT’s requestor interface allows for selection of workers by country, the country on file with Amazon sometimes does not match the country from which the worker’s IP address originates. This occurred for 2 participants in the present data. Berinsky et al. found the same occurrence in 5.8 % of their large sample of U.S. resident participants recruited through AMT and speculated that these participants might be traveling at the time. Rand

Table 3 Comparison of effects found by Lupker et al. (2008, Experiment 1b) and effects found in present data

Subject analysis (F_1)										
Factor	Response times									
	Lupker					Present				
	$F(1, 16)$	p	ω_p^2	ω_p^2 95 % CI		$F(1, 26)$	p	ω_p^2	ω_p^2 95 % CI	
				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>
Alteration type	18.66	<.001	0.495	0.142	0.694	118.26	<.001	0.807	0.640	0.873
Letter type	5.95	.027	0.216	0.000	0.507	19.95	<.001	0.404	0.132	0.602
Alteration \times letter	7.84	.013	0.275	0.015	0.550	14.53	<.001	0.326	0.076	0.545
Error rates										
Factor	Error rates									
	Lupker					Present				
	$F(1, 16)$	p	ω_p^2	ω_p^2 95 % CI		$F(1, 26)$	p	ω_p^2	ω_p^2 95 % CI	
				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>
Alteration type	12.62	.003	0.392	0.068	0.629	25.91	<.001	0.471	0.191	0.649
Letter type	17.28	<.001	0.475	0.125	0.681	9.60	.005	0.235	0.029	0.475
Alteration \times letter	11.76	.003	0.374	0.057	0.617	8.02	.009	0.200	0.016	0.446
Item analysis (F_2)										
Factor	Response times									
	Lupker					Present				
	$F(1, 75)$	p	ω_p^2	ω_p^2 95 % CI		$F(1, 75)$	p	ω_p^2	ω_p^2 95 % CI	
				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>
Alteration type	58.04	<.001	0.760	0.481	0.855	64.71	<.001	0.695	0.463	0.799
Letter type	6.25	.015	0.226	0.001	0.515	10.09	.002	0.245	0.033	0.483
Alteration \times letter	9.26	.003	0.315	0.029	0.577	11.71	.001	0.277	0.048	0.508
Factor	Error rates									
	Lupker					Present				
	$F(1, 75)$	p	ω_p^2	ω_p^2 95 % CI		$F(1, 75)$	p	ω_p^2	ω_p^2 95 % CI	
				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>
Alteration type	46.11	<.001	0.715	0.406	0.827	85.86	<.001	0.752	0.550	0.837
Letter type	14.28	<.001	0.425	0.088	0.650	6.78	.011	0.171	0.007	0.420
Alteration \times letter	24.42	<.001	0.565	0.210	0.737	8.57	.005	0.213	0.020	0.456

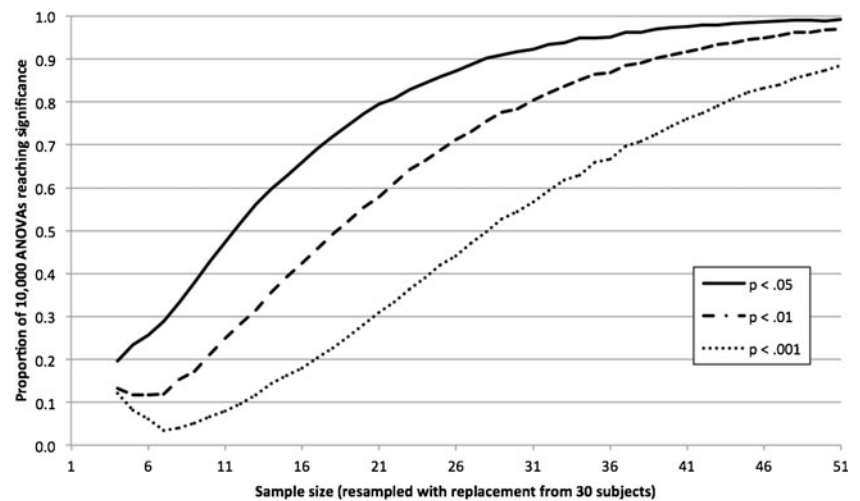
(2012) found that 97.2 % of AMT workers will self-report a country of residence matching that of their IP address. Such discrepancies between the countries of residence in Amazon's records, the countries of origin of a worker's IP addresses, and self-reported countries of residence could be important in interpreting results.

In the speeded forced choice tasks used here, it was easy to determine which participants understood and followed the rules. Several participants responded to the flankers rather than the targets, but their error rates would have been very low had responding to flankers been the goal; they were not simply responding randomly or carelessly. In the

Table 4 Comparison of RTs and rates found by Lupker, et al. (2008, Experiment 1b) and those in present data

Nonword type		Transposed letters		Replaced letters		Transposed-letters effect (TL-RL)	
		Lupker	Present	Lupker	Present	Lupker	Present
		<i>M</i>	<i>M (SD)</i>	<i>M</i>	<i>M (SD)</i>	<i>M</i>	<i>M (SD)</i>
			95 % CI		95 % CI		95 % CI
		<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>
Consonant–consonant		969	975 (146)	868	834 (140)	101	141 (70)
Vowel–vowel		916	903 (156)	867	825 (142)	49	78 (72)
			1,026		887		114
			962		878		51
							105
Error rates (%)							
Nonword type		Transposed letters		Replaced letters		Transposed-letter effect (TL-RL)	
		Lupker	Present	Lupker	Present	Lupker	Present
		<i>M</i>	<i>M (SD)</i>	<i>M</i>	<i>M (SD)</i>	<i>M</i>	<i>M (SD)</i>
			95 % CI		95 % CI		95 % CI
		<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>
Consonant–consonant		17.5	16.5 (13.7)	1.5	2.3 (5.0)	16.0	14.2 (13.1)
Vowel–vowel		7.0	10 (17.9)	3.0	2.7 (5.1)	4.0	7.2 (14.5)
			11.4		0.4		9.4
			14.8		0.8		1.8
					4.6		12.7

Fig. 5 Statistical power of two-way interaction as a function of sample size and significance criterion



participants who did respond to the targets, error rates on both tasks used here were lower than those in the studies that were replicated. In experimental designs with more complicated instructions, requiring participants to pass a comprehension quiz before they can move from the instructions to the experiment can improve data quality (McDonnell, Domingo, & Gureckis, 2012).

In survey research on AMT, strategies of assessing participants' understanding and diligence have been well documented. Goodman et al. (in press) used two methods to assess AMT workers' attention to instructions. One method was the use of instructional manipulation check tests: multiple-choice questions that participants are likely to incorrectly answer if they do not fully read them. The second method was the Cognitive Reflection Test, a set of simple math questions that participants are likely to get wrong if they respond with their first notion rather than actually calculating the answers. They found that a student sample outperformed AMT workers but that AMT workers and a community sample performed similarly. On the other hand, in 2010, Paolacci, Changler, and Ipeirotis administered surveys with embedded catch questions to AMT workers, students, and online discussion board members. AMT workers were more likely than the other groups to correctly answer the catch questions.

The fact that 100 participants were used in this experiment is not meant to imply that such a large sample is necessary to detect an effect. On the contrary, as can be seen in Fig. 4, the evidence suggests that this was not the case. In fact, it shows that, in order to reach significance at an alpha of less than .05 and a beta of less than .05, only 15 randomly selected participants were actually needed. The robustness of the findings with smaller sample sizes is remarkable, considering the diversity of AMT workers as compared with undergraduate student samples, the variety of physical environments in which AMT workers participated, and the differences in the computer configurations of the AMT workers.

Even if a small sample is sufficient for a conventional level of power, a larger sample might be a more reasonable choice when conducting research on AMT. The convention of a power level of .80 ($\beta = .20$) was proposed by Cohen (1988) as a way of balancing the cost of failing to find an effect that does exist with the cost of collecting sufficient behavioral data to find the effect. However, Cohen recommended adopting this power level as a default only when investigators have no other basis for setting the desired power level. The factors involved in research using AMT provide a basis for reconsideration of this balance. In the lab, setting up experiments like those used here is simple, but running 100 participants imposes demands on equipment availability and experimenter time. In contrast, translating an in-lab experiment to an online experiment requires an initial investment of time, while the marginal cost of running each participant is low. Once the experiment is online, the price of participant recruitment is low (e.g., Buhrmester et al., 2011), the rate of data collection can be high (20 participants per hour in this study), and participants provide their own equipment. In Fig. 4, consider the effect of increasing the sample size from 12 to 21 participants. The increase in sample size allows for a shift from an 18 % type II error rate down to 1 %, for an additional \$6.30 in participant payments.

Conclusion

The aim of this study was to validate the effectiveness of the paired use of AMT and Adobe Flash in recruiting participants and conducting mental chronometry experiments. Our results demonstrate that the performance of the pair is sufficient to detect patterns of results previously established in lab-based flanker and lexical decision experiments. Using larger sample sizes than in the lab may not be necessary, but it makes practical sense. When limitations are taken into account in the design of an experiment, the benefits of using

AMT for experimentation beyond survey-type designs can be realized.

Author Note Travis Simcox, Department of Psychology, University of Pittsburgh; The Center for the Neural Basis of Cognition, Pittsburgh; Learning Research and Development Center, University of Pittsburgh. Julie A. Fiez, Department of Psychology, University of Pittsburgh; The Center for Neuroscience, University of Pittsburgh; The Center for the Neural Basis of Cognition, Pittsburgh; Learning Research and Development Center, University of Pittsburgh.

This research was supported by NIH R01 HD060388 and NSF 0815945.

References

- Adobe Developer Connection. (2012). Flash Platform Developer Center. <http://www.adobe.com/devnet/flashplatform.html> Retrieved 2012-08-11.
- AMT FAQ. (2012). https://requester.mturk.com/help/faq#examples_violations Retrieved 2012-08-11.
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*, 800–813.
- Berinsky, A. J., Huber, & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368.
- Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Social Science Computer Review*, *30*, 350–357.
- Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.
- Fowler, R. L. (1985). Point estimates and confidence intervals in measures of association. *Quantitative Methods in Psychology*, *98*, 160–165.
- Goodman, J. K., Cryder, C. E., & Cheema, A. (in press). Data collection in a flat world: Strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making*.
- Grossman, G. & Huang, E. (2009). *ActionScript 3.0 Overview*. Adobe systems incorporated. Retrieved 2012-08-11 from http://www.adobe.com/devnet/actionscript/articles/actionscript3_overview.html
- Halberda, J., Ly, R., Wilmer, J., Naiman, D., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive internet-based sample. *Proceedings of the National Academy of Sciences*, *109*, 11116–11120.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavior and Brain Sciences*. Retrieved from: <http://www2.psych.ubc.ca/~henrich/pdfs/WeirdPeople.pdf>
- Hewson, C. M., Laurent, D., & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments, & Computers*, *28*, 186–191.
- Houben, K., & Wiers, R. W. (2008). Measuring implicit alcohol associations via the Internet: Validation of Web-based implicit association tests. *Behavior Research Methods*, *40*, 1134–1143.
- Ipeirotis, P. (2010). Demographics of Mechanical Turk. CeDER Working Papers, CeDER-10-01, New York University, Stern School of Business. Retrieved Aug 2012 from: <http://hdl.handle.net/2451/29585>
- Lupker, S. J., Perea, M., & Davis, C. M. (2008). Transposed-letter effects: Consonants, vowels and letter frequency. *Language & Cognitive Processes*, *23*, 93–116.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1–23.
- Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. In *Proc. Interspeech*, Portland, OR, USA.
- McDonnell, J., Domingo, D., & Gureckis, T. (2012). *Is Mechanical Turk the future of cognitive science research?* Retrieved Aug 2012 from <http://gureckislab.org/blog/?p=1297>
- Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology*, *26*, 3–67.
- Meyer, D. E., & Roger, W. S. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.
- Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in Apple Macintosh computers. *Behavior Research Methods*, *43*, 353–362.
- Nieuwenhuis, S., Stins, J. F., Posthuma, D., Polderman, T. J., Boomsma, D. I., & de Geus, E. J. (2006). Accounting for sequential trial effects in the flanker task: Conflict adaptation or associative priming? *Memory & Cognition*, *34*, 1260–1272.
- Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., et al. (2010). Putting brain training to the test. *Nature*, *465*, 775–778.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.
- Paré, D. E., & Cree, G. S. (2009). Web-based image norming: How do object familiarity and visual complexity ratings compare when collected in-lab versus online? *Behavior Research Methods*, *41*, 699–704.
- Plant, R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*, 598–614.
- R Developmental Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>
- Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *222*, 172–179.
- Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, *39*, 365–370.
- Schmidt, W. C. (2001). Presentation accuracy of Web animation methods. *Behavior Research Methods, Instruments, & Computers*, *33*, 187–200.
- Simcox, T. (2012). [Compilation of user agent strings from Amazon Mechanical Turk workers]. Unpublished raw data.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605–632.
- The Mechanical Turk Blog. (2012, August 28). Improving quality with qualifications – tips for API requesters [Web log post]. Retrieved from: <http://mechanicalturk.typepad.com/blog/2012/08/requesters-consistently-tell-us-that-using-qualifications-is-one-of-the-most-effective-strategies-for-optimizing-the-quality.html>
- Ulrich, R., & Giray, M. (1989). The resolution of clocks: Effects on reaction time measurement-Good news for bad clocks. *British Journal of Mathematical and Statistical Psychology*, *42*, 1–12.
- Woltman, G. (2012). *Prime95*. Retrieved from: <http://www.mersenne.org/freesoft/>
- Yu, C. H. (2003). Resampling methods: Concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, *8*. Retrieved from <http://PAREonline.net/getvn.asp?v=8&n=19>