Behav Res
DOI 10.3758/s13428-015-0599-7

Conducting spoken word recognition research online: Validation and a new timing method

Joseph Slote[1] · **Julia F. Strand**[1]

Abstract Models of spoken word recognition typically make predictions that are then tested in the laboratory against the word recognition scores of human subjects (e.g., Luce & Pisoni *Ear and Hearing*, *19*, 1–36, 1998). Unfortunately, laboratory collection of large sets of word recognition data can be costly and time-consuming. Due to the numerous advantages of online research in speed, cost, and participant diversity, some labs have begun to explore the use of online platforms such as Amazon's Mechanical Turk (AMT) to source participation and collect data (Buhrmester, Kwang, & Gosling *Perspectives on Psychological Science*, *6*, 3–5, 2011). Many classic findings in cognitive psychology have been successfully replicated online, including the Stroop effect, task-switching costs, and Simon and flanker interference (Crump, McDonnell, & Gureckis *PLoS ONE*, *8*, e57410, 2013). However, tasks requiring auditory stimulus delivery have not typically made use of AMT. In the present study, we evaluated the use of AMT for collecting spoken word identification and auditory lexical decision data. Although online users were faster and less accurate than participants in the lab, the results revealed strong correlations between the online and laboratory measures for both word identification accuracy and lexical decision speed. In addition, the scores obtained in the lab and online were equivalently correlated with factors that have been well established to predict word recognition, including word frequency and phonological neighborhood density. We also present and analyze a method for precise auditory reaction timing that is novel to behavioral research. Taken together, these findings suggest that AMT can be a viable alternative to the traditional laboratory setting as a source of participation for some spoken word recognition research.

Keywords Spoken word recognition · Online data collection · Reaction time

Collecting data for behavioral research can be time consuming and expensive for researchers, and tedious for participants. In order to make the process tractable, researchers are often forced to limit their trial, participant, or stimulus counts. Because of these and other disadvantages of laboratory-based data collection, some researchers have turned to the Internet as an alternative source of participation. Online experimentation has many attractive qualities. First, participants tend to be more diverse than university subject pools and are willing to participate for less compensation, thanks to their ability to participate at the time and place of their choosing (Paolacci, Chandler, & Ipeirotis, 2010). Data from many participants can be collected more quickly than in traditional laboratory settings and during periods when recruiting undergraduate participants is difficult, such as between terms (Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2011). In addition, given that online users never interact with an experimenter and have no preconceptions about the kinds of studies being done in a particular research lab, online data collection may help avoid experimenter bias or effects of participant expectation. Finally, because the experiment runs in each participant's browser, participation can be highly parallelized, the instructions are identical for each participant, and the experiment procedure can be easily shared with other researchers in the form of source code.

Despite the advantages of Web experimentation, two major factors have historically limited its adoption. First, it was a

✉ Julia F. Strand
jstrand@carleton.edu

[1] Department of Psychology, Carleton College, Northfield, MN 55057, USA

Published online: 19 May 2015

challenge to recruit participants at a sufficient rate to warrant the study's presence online. Second, a lack of control over who participated and the environment in which the tasks were completed raised concerns over the validity of data collected online. Not only are researchers absent from the experiment environment to ensure that the participant is taking the study seriously, but unlike in a carefully controlled laboratory setting, one cannot guarantee the technological ability of the participants' computer systems. The feasibility of multimedia stimuli or millisecond-resolution timing in online research has been demonstrated only recently, and concerns about performance differences across participant devices linger (see Reimers & Stewart, 2014, for an in-depth discussion of the development in each of these areas).

Amazon Mechanical Turk (AMT), an online labor market for short tasks, has proven to be a worthy solution to the challenge of participant recruitment (see Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2011, for an introduction to behavioral research using AMT). With very little extra effort or overhead cost, behavioral researchers have been able to achieve very high participation rates in considerably shorter times than would be possible in traditional laboratory settings. The service includes a built-in feature to prevent duplicate participation, and researchers are able to reject (prior to compensation) responses that appear to be incomplete or incompatible with the instructions. Furthermore, Gureckis and colleagues (McDonnell et al., 2012) have developed an open-source and ever-improving framework called psiTurk that provides a common starting point for behavioral psychology experiments on AMT. PsiTurk facilitates behavioral research by streamlining compensation management, data storage, and experiment development and deployment.

The second concern, regarding the potentially adverse effects of participant and environmental variability, has been addressed for a number of standard behavioral tasks. Although some studies have demonstrated differences between the data collected online and in the lab (see Crump et al., 2013), many phenomena, including the Stroop, flanker, subliminal-priming, and Posner cueing tasks (Crump et al., 2013), as well as framing and representativeness heuristics in decision making (Paolacci et al., 2010), have been replicated online. Notably, these replications include a wide range of behavioral tasks, including problem solving and learning, as well as those that require precise millisecond measurement and control. These validation studies suggest that the practical advantages of using AMT do not come at the cost of compromised data.

However, little work to date has evaluated the use of online data collection for studies that require listening to auditory stimuli, and even less has examined spoken word recognition online (but see Cooke, Barker, Garcia Lecumberri, & Wasilewski, 2011). Conducting auditory research online poses several challenges in addition to those faced by studies that employ visual stimuli alone. First, although Web technologies provide tools to effectively standardize the presentation of simple visual stimuli (such as individual words or images), researchers have significantly less control over the quality and amplitude of audio stimuli, and, historically, the precision of the onset time of the stimulus. When conducting research on spoken word recognition in the lab, researchers carefully determine a signal-to-noise ratio and overall amplitude at which to present stimuli in order to avoid floor and ceiling effects. Stimuli are typically presented to participants via high-quality headphones or speakers in a sound-attenuating chamber with no visual distractions. On the other hand, not only will audio hardware (i.e., speakers or headphones) vary among online users, but AMT users also have control over the volume at which their computers play sounds, making it impossible to ensure that auditory stimuli are presented at a consistent amplitude across participants. In addition, researchers have no control over the auditory environment in which AMT users complete the task. It is therefore likely that some participants will be listening in settings that include background noise at levels above what would be acceptable in the laboratory.

Despite these concerns, data that have been collected using auditory stimuli online thus far have demonstrated some key similarities with data collected in the lab. Cooke et al. (2011) found that participants in the lab and participants online are similarly affected by changes in signal-to-noise ratio and the type of masker noise (i.e., multitalker babble, speech shaped noise, etc.). Participants in the lab and on AMT also showed similarities in rating the intelligibility of different speech types, including infant-directed speech, computer-directed speech, and other types (Mayo, Aubanel, & Cooke, 2012). However, differences between data sets obtained in the lab and online have also been identified. For example, online users show some discrepancies in the patterns of speech sounds that they confuse (Cooke, Barker, Lecumberri, & Wasilewski, 2013), and word recognition scores are consistently lower online than in the lab for both natural (Cooke et al., 2013) and synthetic (Wolters, Isaac, & Renals, 2010) speech. Therefore, additional research is needed to evaluate the conditions under which spoken word recognition data collected online are comparable to lab-collected data.

Online auditory experimentation is further complicated when the experimental task involves measurements of participants' reaction times (RTs). Because of the computational load that decoding, buffering, and playing audio requires, modern computer architecture offloads the task to a separate hardware component, the soundcard. This device has its own internal clock and, because Web technology in general has very limited programmatic access to computer hardware, it has traditionally been difficult either to obtain the time that an audio clip began or to align the beginning of an audio clip with a specific time set by the main processor. It is reasonable,

therefore, to doubt that RTs in response to auditory stimuli collected from an online platform such as AMT could be accurate enough to expose subtle linguistic effects. For the present study, we took a multipronged approach to addressing this concern by using a timing method provided by recent developments in Web technology.

In Experiment 1, we conducted two standard spoken word recognition tasks both in the laboratory and online, and then compared the results from the two settings. In addition, we evaluated how these word recognition scores correlated with lexical variables that have been established as consistent predictors of recognition accuracy. In Experiment 2, we verified the performance of the timing method used in Experiment 1 directly, by comparing it to a naive timing solution.

Experiment 1

The most commonly used tasks in research on spoken word recognition are word identification in noise (Pisoni, 1996) and auditory lexical decision (Goldinger, 1996). Word identification in noise (ID) tasks typically involve presenting participants with individual words in a background of masking noise, such as white noise or multitalker babble, and asking them to try to identify the word. In an auditory lexical decision (ALD) task, participants are presented with words and phonotactically legal nonwords and are asked to determine as quickly and accurately as possible whether the stimulus that they heard formed a real word, and to respond by pressing a button.

From a theoretical standpoint, much research on spoken word recognition has sought to describe the process by which a stimulus word is disambiguated from all other words in the mental lexicon (see Dahan & Magnuson, 2006, and Weber & Scharenborg, 2012, for reviews). Although models of word recognition differ in implementation, they do include mechanisms to explain why some words are identified more quickly and accurately than others (cf. Luce, Goldinger, Auer, & Vitevitch, 2000; Luce & Pisoni, 1998; McClelland, Elman, & Diego, 1986). The most well-established factor that predicts word identification accuracy is the frequency with which the word occurs in language: Common words are identified more quickly and accurately than rare words (Brysbaert & New, 2009; Savin, 1963). A second factor that robustly predicts word recognition scores is the perceptual similarity of the stimulus word to other words in the mental lexicon. Models of recognition assume that stimulus input in the form of the acoustic signal activates multiple lexical candidates (often called "neighbors") in memory, and that these candidates then compete with one another for recognition. Therefore, due to this lexical competition, words with many neighbors are identified more slowly and less accurately than words that are more distinct (Luce & Pisoni, 1998; Vitevitch & Luce, 1998). In addition, frequency also appears to modulate the effects of lexical competition; words that tend to have more high-frequency neighbors are recognized more slowly and less accurately than words with low-frequency neighbors (Luce & Pisoni, 1998). Both the ID and ALD tasks are assumed to be influenced by the organization of the mental lexicon and are sensitive to word frequency and lexical competition effects.

The goals of Experiment 1 were twofold. First, we sought to evaluate whether ID and ALD data collected using AMT are comparable to data collected in the laboratory. Second, we assessed whether data collected using AMT are affected by lexical variables at rates comparable to those of data collected in the laboratory.

Method

Participants

Laboratory The participants were native English speakers ($N = 53$ in the ID task, $N = 51$ in the ALD task) with self-reported normal hearing and normal or corrected-to-normal vision, who were recruited from the Carleton College undergraduate student body. Testing took approximately 30 min, and participants were awarded $5 for their time. Carleton College's Institutional Review Board approved the research procedures.

AMT The experiment was programmed in JavaScript using the psiTurk experiment platform (McDonnell et al., 2012). Online data were collected between the dates of July 30 and August 6, 2014. Workers on the AMT residing in the United States were presented with an advertisement for the study that listed various personal, environmental, and technical requirements: that they have normal hearing, be in a quiet environment, and use a modern Web browser. All workers self-reported being native English speakers or reported speaking English most of the time. Testing took approximately 30 min and participants were awarded $2.50 for their time. Different groups of participants completed the ALD and ID tasks ($n = 100$ for each) and were randomly assigned by psiTurk's condition balancing algorithm. An additional 76 (ID: $n = 34$; ALD: $n = 42$) participants began the study but failed to complete it for unknown reasons, rendering a completion rate of 72 %.[1] Carleton College's Institutional Review Board approved the research procedures.

[1] The data on completion rates for psychological studies on AMT are not widely available, but Crump et al. (2013) reported a completion rate of 81 %. Future work should further explore the factors that influence attrition in online studies.

Stimuli

The stimuli for the ID and ALD tasks included 400 consonant–vowel–consonant (CVC) words, selected to ensure a range of values of lexical variables, including frequency and lexical neighborhood size. The ALD task also included 400 phonotactically legal CVC nonwords (e.g., "dak," "lin"). Speech stimuli were recorded at 16 bits, 44100 Hz using a Shure KSM-32 microphone with a pop filter, by a female speaker with a standard Midwestern accent, and were equalized on total root mean squared intensity (RMS) using Adobe Audition, version 5.0.2. In the ID task, speech stimuli were presented in a background of six-talker babble (signal-to-noise ratio = 0). The ALD stimuli were presented without background noise. In the laboratory, both the ALD and ID stimuli were presented at approximately 65 dB through Sennheiser HD-280 PRO headphones.

Procedure

In both the ID and ALD tasks, participants in the laboratory were seated in a quiet room a comfortable distance from an iMac computer running the Cedrus Superlab 5.0 stimulus presentation software. In the ID task, the lab and AMT participants were presented with isolated auditory stimuli in a randomized order. They then typed their response into a white text box with large, black font displayed in the middle of a gray screen. Participants were encouraged to guess when they were unsure. After entering each response, a 1-s intertrial interval elapsed before the next word was presented. A short practice session consisting of five additional CVC words preceded the experiment. Participants completed the full task in a single block without breaks.

In the ALD task, the lab and AMT participants were presented with a blank gray screen and heard the stimuli in a randomized order. Lab participants responded with a Cedrus Response Pad (RB730), whereas online participants used the Tab and Backslash keys of their keyboards to indicate "nonword" and "word," respectively. The subsequent trial began after a 250-ms postresponse interstimulus interval. The on-screen display was identical in the lab and online, except for the presence of a keyboard legend in the AMT version. A short practice session consisting of two CVC words and two CVC nonwords preceded the experiment.

The procedures in the lab and online were designed to be as similar as possible. However, some extra precautions were included in the online version, as an attempt to mitigate distraction and verify technical sufficiency. The AMT users were first presented with an audio CAPTCHA via Google's reCAPTCHA service (Google, 2014) that required them to transcribe several numbers in a challenging listening situation. This was done for multiple reasons: to dissuade users from using computer scripts ("bots") to take the study, to verify sufficient hearing ability, and to ensure that the participant's audio equipment was functioning properly and was set at an amplitude appropriate for the task.

Participants were also required to put their browser in full-screen mode in order to mitigate distraction from other software. If a participant exited full-screen mode during the experiment, the study was paused and input was blocked until the participant reentered full-screen mode. The participants who paused the experiment in this way were allowed to continue, and their data were treated identically in the subsequent analysis to the data from all other participants; we do not believe this allowance affected the results in a systematic way, because the mean and standard deviations for the time required to complete the test trials of the experiment were very similar for data in the lab (ID, $M = 18$ min, $SD = 6$ min; ALD, $M = 19$ min, $SD = 4$ min) and on AMT (ID, $M = 20$ min, $SD = 7$ min; ALD, $M = 19$ min, $SD = 6$ min).

In both the ID and ALD tasks, the audio was preloaded, buffered, and presented using the Web Audio API (Adenot & Wilson, 2015). The total RMS amplitudes of the audio stimuli were adjusted to match the level of the samples used by the reCAPTCHA service. Online RTs were collected via the *currentTime* property of the *AudioContext* interface (see Exp. 2 for implementation details). Source code for the AMT experiment is available at http://go.carleton.edu/StrandLab.

Results and discussion

Word identification

Prior to compensation, the data of the AMT participants who responded with less than 10 % accuracy on the ID task were manually checked for responses incompatible with the task instructions (e.g., empty strings or nonsense words). This resulted in the rejection of two online participants' work. Both the in-lab and online responses were then were hand-checked for obvious typographical errors. Entries were corrected if they included extraneous punctuation (e.g., "fit/"), were phonologically identical to the target (e.g., "sighed" and "side"), and when the entry did not represent a real word but differed from the target by one letter (e.g., "calfr" to "calf"). These corrections represented approximately 1 % of the responses in both the lab data and the AMT data. Word identification accuracy was then calculated for each of the target words in both the lab data and the AMT data.

Words were identified significantly more accurately in the lab than on AMT, $t(399) = 21.81$, $p < .001$, Cohen's $d = 0.54$, with lab users scoring an average of 14 % higher than AMT users (see Fig. 1).

Although overall accuracy was higher in the lab than on AMT, there was a strong correlation between the word identification accuracy scores obtained in the lab and online, $r =$

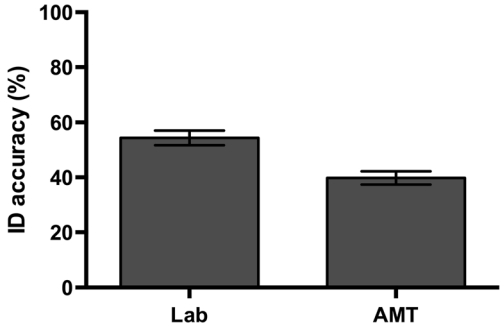

Fig. 1 Identification accuracy. Error bars represent 95 % confidence intervals

.87, $p < .001$ (see Fig. 2), indicating that words that were more difficult for participants to identify in the lab were also more difficult for online users. This correlation is similar in magnitude to the split-half reliabilities of the ID task both in the lab ($r = .90$, $p < .01$) and on AMT ($r = .94$, $p < .001$), indicating that some of the deviation between the scores in the lab and online was simply a function of noise in the replication process, rather than a systematic difference between in-lab and online data collection.

Auditory lexical decision task

Similarly to the ID task, the data of AMT participants who responded with less than 10 % accuracy on the ALD task were manually screened for responses incompatible with the task instructions (e.g., answering "nonword" for every stimulus). No participants responded in an obviously incompatible way. In-lab and online participants' individual ALD responses that were longer than 2,000 ms were excluded. These made up fewer than 3 % of all ALD responses. The average latencies for all correct responses to the word stimuli were then calculated for each stimulus word. Ten words had accuracy rates less than 40 % (three standard deviations below the mean) in the lab data and/or the AMT data. Given that these stimuli would include a very small number of correct responses from which to draw latency data, the ALD analysis was conducted on the remaining 390 words. To account for the influence of

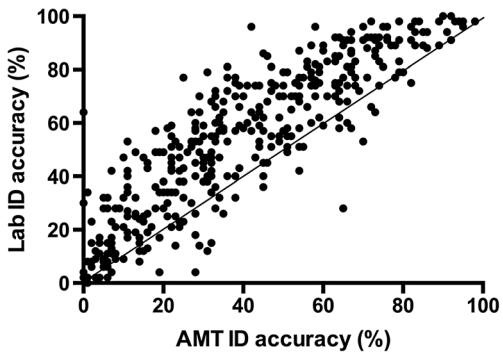

Fig. 2 Identification accuracy for each stimulus word, in the lab and on Amazon Mechanical Turk (AMT). The line represents $x = y$

word duration on RTs, latencies were measured from the offset of the stimulus word. Studies that employ longer-length stimuli (which may be identified prior to offset) should consider measuring the latency from word onset and entering the stimulus length as a covariate. Given the short length of the materials in the present study, the major findings did not differ if the latency was measured from word onset rather than offset. As compared to the data collected in the laboratory, the responses from AMT were 63 ms faster, $t(389) = 26.61$, $p < .001$, Cohen's $d = 0.75$, and 5 % less accurate, $t(389) = 17.87$, $p < .001$, Cohen's $d = 0.54$ (see Fig. 3).

In parallel to the findings of the ID task, there was a strong correlation between the ALD latency data collected in the lab and on AMT, $r = .86$, $p < .001$, and between the ALD accuracy data collected in the lab and on AMT, $r = .82$, $p < .001$; see Fig. 4. Again, these results were similar to the split-half reliabilities of the lab data (RT data, $r = .84$, $p < .001$; accuracy data, $r = .85$, $p < .001$) and the AMT data (RT data, $r = .83$, $p < .001$; accuracy data, $r = .89$, $p < .001$).

Links with lexical variables

In addition to evaluating the reliability of the measures collected online and in the laboratory, in the present study we also sought to assess whether previously used lexical variables explained similar amounts of variance in the data collected from the lab and online. These variables were selected on the basis of their well-established role in predicting word identification accuracy and latency in other studies, and they included word frequency (Brysbaert & New, 2009), age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), familiarity (Connine, Mullennix, Shernoff, & Yelen, 1990), neighborhood size (Luce & Pisoni, 1998), neighborhood frequency (Luce & Pisoni, 1998), phi-square density (Strand, 2014), and phonotactic probability (Vitevitch, Luce, Pisoni, & Auer, 1999).

Word frequency values were obtained from the data set of Brysbaert and New (2009), which calculated frequency counts from spoken television and film subtitles. Age-of-acquisition data were obtained from an existing data set (Kuperman et al., 2012) that assessed the ages at which individuals first learn words. Words that are learned younger tend to be recognized more easily than those that are learned later (Turner, Valentine, & Ellis, 1998). Familiarity values were obtained from the Hoosier Mental Lexicon (Sommers, 2002); familiarity facilitates word recognition (Connine et al., 1990).

Lexical competition has most commonly been quantified by defining "neighbors" as words that may be formed by a single, position-specific phoneme addition, deletion, or substitution (see Luce, Pisoni, & Goldinger, 1990). Values for the number of neighbors were obtained from an existing database (Balota et al., 2007). We also calculated the average frequency (Brysbaert & New, 2009) of a word's neighborhood, given

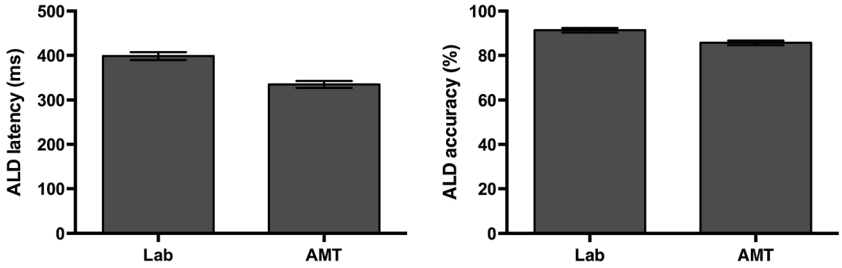

Fig. 3 Average latencies and accuracies for words in the auditory lexical decision (ALD) task in the lab and on Amazon Mechanical Turk (AMT). Error bars represent 95 % confidence intervals

prior work demonstrating that words with higher-frequency neighbors tend to be identified less accurately than those with lower-frequency neighbors (Luce & Pisoni, 1998). Lexical competition has also been quantified on a continuous scale, by assessing the perceptual similarity of a target word to every other word in the lexicon, using the probabilities that the two words' segments will be confused on a forced choice phoneme identification task (Luce & Pisoni, 1998; Strand, 2014). One such continuous measure of lexical competition, phi-square density, quantifies the amount of lexical competition for each stimulus word by evaluating the expected confusability of each word with all other words in a lexicon (see Strand & Sommers, 2011, and Strand, 2014, for methodological and computational details). We adopted phi-square density as an additional measure of lexical competition, and values were obtained from the Phi-lex database (Strand, 2014). Although categorical (neighbor-based) and continuous (e.g., phi-square density) measures of lexical competition are correlated with one another and account for variance in word recognition accuracy, phi-square density accounts for significantly more unique variance in spoken word recognition accuracy than do neighbor-based approaches (Strand & Sommers, 2011; Strand, 2014). Both measures are included here to more rigorously evaluate the similarity of lab data and AMT data by using multiple measures of lexical competition. Finally, we also obtained measures of phonotactic probability, a metric of the frequency of occurrence of a given words' segments (Vitevitch & Luce, 2004). Words with high-probability segments tend to be recognized more quickly than those with low-probability segments (Vitevitch et al., 1999).

The influences of the seven lexical variables were evaluated for both the lab and AMT measures of ID accuracy and ALD latency.[2] The magnitudes of the correlations of the lexical variables with the lab and AMT measures are quite similar (see Tables 1 and 2). In line with prior research, higher-frequency words were identified more quickly and accurately than lower-frequency words. Age of acquisition predicted word identification accuracy and ALD latency in both the lab and on AMT, with facilitation for words learned younger. The correlation with word familiarity only reached significance for the AMT ID data. Words with more lexical competition (as measured by number of neighbors or phi-square density) were identified more slowly and less accurately. Words with neighbors that tend to be high-frequency were recognized moderately more accurately both in the lab and on AMT, but did not influence RTs. This finding is somewhat surprising, because neighbor frequency tends to be detrimental to identification accuracy (Luce & Pisoni, 1998). However, when controlling for target word frequency, the relationship between neighbor frequency and accuracy disappears (*p*s > .31 for both comparisons), suggesting that the correlation between neighbor frequency and identification accuracy is due to collinearity between target word frequency and neighbor frequency. Phonotactic probability was significantly correlated with ALD latencies in both the lab and AMT data, although not in the ID data. Fisher *r*-to-*z* transformations revealed no significant differences in the magnitudes of the correlations between lab- and AMT-derived measures and the lexical variables.

Given the degree of multicollinearity between multiple lexical variables (e.g., frequency and age of acquisition or phi-square density and number of neighbors), we also conducted a series of multiple regressions to evaluate the unique variance explained by each predictor variable in the lab and the AMT data. The seven measures of lexical competition were entered in a stepwise multiple regression, which followed a forward selection approach but also evaluated whether the removal of a predictor improved the model at each step (Field, 2009). No previously selected variables were removed in any of our analyses, so the results were identical to a forward-selection approach (see Tables 3 and 4). Given the finding that lexical competition effects may be moderated by frequency (Goh, Suárez, Yap, & Tan, 2009; Luce & Pisoni, 1998), we also included a term for the Neighborhood Size × Frequency interaction, but this failed to account for significant unique variance in either the lab data or the AMT data.

A parallel analysis was conducted for the ALD data, using the same seven lexical variables. Only three predicted significant unique variance in ALD latencies: frequency, phi-square

[2] Due to the high values and low variability of ALD accuracy, these data were not examined further.

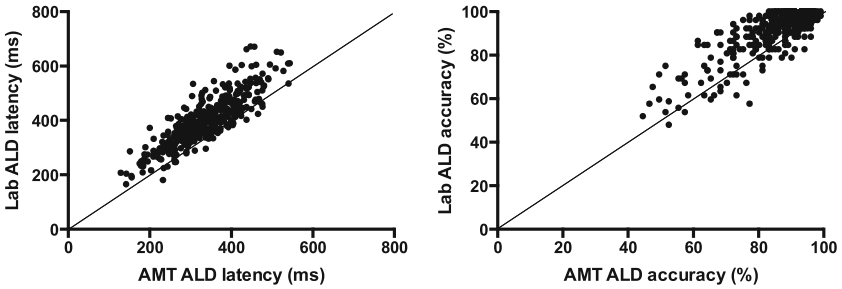

Fig. 4 Auditory lexical decision (ALD) latency and accuracy for each stimulus word, in the lab and on Amazon Mechanical Turk (AMT). The lines represent $x = y$

density, and neighborhood size (see Table 4). The remaining four variables and the Neighborhood Size × Frequency interaction measure failed to account for significant variance.

As in the ID data, the ALD regression analyses demonstrated strong consistencies between the data collected in the lab and online. However, these comparisons were being made on different sample sizes, since the AMT sample had nearly double the participants that the lab sample did. To evaluate whether these different sample sizes influenced our results, we also completed the regressions above using a random sample of the AMT participants to match the size of the lab sample. The major outcomes did not change, indicating that the larger size of the AMT sample was not responsible for the similarity with the lab data. However, future studies that are concerned about the possibility of greater variability in the AMT sample should evaluate whether larger samples are necessary for sufficient power.

An additional analysis that researchers have used in studies on spoken word recognition is to compare the accuracies and latencies for words that that vary in lexical "difficulty" (Kaiser, 2003; Luce & Pisoni, 1998; Sommers, 1996; Sommers & Danielson, 1999). "Easy" words are those that are high in frequency and have relatively few neighbors that tend to be low-frequency. "Hard" words are low-frequency words with many high-frequency neighbors. In the present data set, easy and hard words were selected as those that were above or below the median value on each characteristic, resulting in 60 easy words and 52 hard words. As compared to hard words, the easy words were higher in frequency, $t(110) = 12.66$, $p < .001$, Cohen's $d = 2.43$, had fewer neighbors, $t(110) = -15.30$, $p < .001$, Cohen's $d = 2.90$, and had lower-frequency neighbors, $t(110) = -4.44$, $p < .001$, Cohen's $d = 0.88$. In the ID task, words were identified more accurately in the lab than online, $F(110, 1) = 117.66$, $MSE = .009$, $p < .001$, $\eta_p^2 = .52$, and easy words were identified more accurately than hard words, $F(110,1) = 18.95$, $MSE = .12$, $p < .001$, $\eta_p^2 = .15$. Critically, the Difficulty × Data Collection Method interaction was not significant, $F(1, 110) = 0.30$, $MSE = .009$, $p = .59$, $\eta_p^2 = .003$, indicating that the influence of lexical difficulty was consistent across the AMT and lab data; see Fig. 5.

A parallel analysis in the ALD data revealed the same pattern. Words were identified more quickly on AMT than in the lab, $F(1, 110) = 162.93$, $MSE = 1{,}298.53$, $p < .001$, $\eta_p^2 = .60$, easy words were identified more quickly than hard words, $F(1, 110) = 39.41$, $MSE = 1{,}219.15$, $p < .001$, $\eta_p^2 = .26$, and there was no interaction between lexical difficulty and data collection method, $F(1, 110) = 1.89$, $MSE = 1{,}298.53$, $p = .17$, $\eta_p^2 = .02$.

Browser and operating system statistics

Participant characteristics such as age or technological ability may influence the hardware and software that they use. Therefore, it is possible that participants who use particular browsers or operating systems may differ systematically in performance. To assess this, we evaluated whether browser and operating system choices influenced the performance on all tasks. The majority of the AMT participants used Windows computers to complete the task (84 %), with 15 % using MacOS and 1 % using Linux. Chrome was the most common Web browser (81 %), with 17 % using Firefox and 2 % using Safari. We observed no systematic differences among the operating system or browser types. That is, ID accuracies, ALD accuracies, and ALD latencies were equivalent across operating systems and browser types ($ps > .17$ for all comparisons).

Table 1 Correlations between lexical variables and identification-in-noise (ID) measures

	Frequency	age of acquisition	Familiarity	Number of Neighbors	Neighbor Frequency	Phi-Square Density	Phonotactic Probability
Lab ID ACC	.30**	–.21**	.05	–.11*	.13**	–.27**	–.01
AMT ID ACC	.34**	–.21**	.12*	–.11*	.09x	–.30**	–.04

AoA, age of acquisition; ACC, accuracy; AMT, Amazon Mechanical Turk. x $p = .08$, * $p < .05$, ** $p < .01$

Table 2 Correlations between lexical variables and auditory lexical decision (ALD) measures

	Frequency	AoA	Familiarity	Number of Neighbors	Neighbor Frequency	Phi-Square Density	Phonotactic Probability
Lab ALD latency	–.38**	.26**	–.02	.15**	.02	.18**	.14**
AMT ALD latency	–.35**	.24**	–.04	.18**	–.03	.20**	.16**

AoA, age of acquisition; AMT, Amazon Mechanical Turk. * $p < .05$, ** $p < .01$

Performance across the tasks

Given the length of the study and the relatively tedious nature of the task, participant fatigue, and therefore impaired performance later in the task, might be a concern, particularly for AMT, on which users might be less motivated. To assess this, we compared the accuracies and latencies on the first half of the tasks to those on the second half. Contrary to the predictions of a fatigue account, performance was higher on the second half of the ID task both in the lab [6 % increase; $t(398) = 8.56$, $p < .001$] and on AMT [3 % increase; $t(398) = 5.90$, $p < .001$]. Latencies in the ALD task were faster in the second half than the first half both in the lab [16-ms decrease; $t(798) = 6.54$, $p < .001$] and on AMT [39-ms decrease; $t(798) = 17.90$, $p < .001$]. Given the well-established effect of talker familiarity on word recognition (Nygaard & Pisoni, 1998), this may be a function of learning the speaking style of the talker, along with gaining familiarity with the task. Given that the stimuli were presented in a random order to each participant, these increases in performance over the course of the task could not systematically influence evaluating the links with lexical variables. Future studies whose results may be influenced by these types of changes in performance across time should consider counterbalancing or randomizing the order in which stimuli or conditions are presented.

Taken together, these results demonstrate robust consistencies between data collected in the laboratory and on AMT. Specifically, we found strong correlations between word identification accuracies and latencies, and similar correlations with lexical variables. Although the relative performance was consistent across the settings, the data revealed significant differences in the magnitudes of accuracy and speed in the lab and online measures. These findings are consistent with prior research showing that AMT users are less accurate overall than lab users (Cooke et al., 2011), but we are the first to show correlations between individual stimulus items in laboratory and AMT data and to demonstrate relationships with lexical variables.

Although the present data cannot explain why AMT users were faster and less accurate than lab users, it is possible that environmental factors and task demands influenced these differences. For instance, the disparity in accuracy may be partially attributable to the overall poorer quality of the listening experience of AMT users. Assuming that the average AMT user was completing the task in a listening context inferior that of a lab user (i.e., noisy background, lower-quality headphones), the overall reductions in accuracy might simply be a function of a more difficult signal-to-noise ratio. The latency differences might be attributable to a contrast in priorities: Undergraduates participating in lab studies may prioritize accuracy over speed, whereas AMT users are likely to be completing the task as quickly as possible in order to move on to the next task and optimize monetary gain. This increase in speed may have come at the cost of lower accuracies in the ALD and ID tasks.[3]

Table 3 Results of a regression predicting word recognition accuracy in the lab and on Amazon Mechanical Turk (AMT) from the lexical variables

	Lab		AMT	
	Beta	R^2 change	Beta	R^2 change
Frequency	.25	.09**	.31	.12**
Phi-square density	–.39	.10**	–.42	.13**
Age of acquisition	–.14	.01*	–.13	.01*
Phonotactic probability	.16	.01*	.14	.01*
Number of neighbors	–.12	.01*	–.10	.01*
Average neighbor frequency	.12	.01*	.07	.00
Familiarity	.05	.00	.13	.01*
Total R^2		.24		.29

* $p < .05$, ** $p < .01$. Familiarity and average neighbor frequency were not selected as significant predictors in the lab and the AMT data, respectively. Betas reflect values at the final step.

Experiment 2

Many behavioral tasks (including the ALD task used in Exp. 1) rely on the ability of the researcher to precisely time participants' responses. This is straightforward in the lab, where it is

[3] As one reviewer pointed out, researchers concerned about this speed–accuracy trade-off may motivate AMT users to prioritize accuracy by imposing an accuracy criterion that participants must reach prior to compensation.

Table 4 Results of a regression predicting auditory lexical decision latencies in the lab and on Amazon Mechanical Turk (AMT) from the lexical variables

	Lab		AMT	
	Beta	R^2 change	Beta	R^2 change
Frequency	–.43	.15**	–.41	.12**
Phi-square density	.21	.06**	.22	.07**
Number of neighbors	.14	.02**	.17	.03**
Total R^2		.23		.22

* $p < .05$, ** $p < .01$

common practice to ask participants to respond via devices such as voice keys or button boxes that have fine temporal resolutions with known tolerances. Online, however, differences in hardware performance, along with varying and unknown amounts of presentation and response lag, may introduce confounding noise into the measurement.

Rather than being able to choose well-suited stimulus presentation and input systems for their experiment, researchers conducting an online study may only influence the accuracy of their timing measurements by carefully programming their experiments on a software platform chosen from a small set of commonly available options (AMT prohibits asking participants to download specialized software to complete a task). Although a variety of such platforms are in use (e.g., Simcox & Fiez, 2014), JavaScript, the Web's native programming language, is becoming increasingly attractive in comparison to plugin alternatives such as Flash or Java. As was noted by Reimers and Stewart (2014), JavaScript is nonproprietary, supported by all modern browsers, and requires no extra software to function (see Crump et al., 2013, for examples of experiments that have used JavaScript on AMT). Meanwhile, the US Department of Homeland Security has recommended that users uninstall Java 7 from their machines due to serious security concerns discovered in 2013, and Adobe has ceased developing Flash for mobile devices.

Despite its appeal, JavaScript is not without its limitations. Because it is a scripting language native to the Web browser environment, the experiment code is transmitted to the user uncompiled. This means that a skilled participant may be able to manipulate the experiment to skip trials, trigger rewards, or create a bot to take the experiment multiple times. Fortunately, AMT makes it possible to programmatically (or manually) check for manipulation prior to compensation, mitigating this risk. Furthermore, any time advantage that would result from writing a script to automate participation is made irrelevant by the ease with which a researcher can prevent duplicate participation. Unless the experiment is hours long, a participant has little incentive to invest the time required to convincingly provide false data instead of simply participating in the study.

Another difficulty associated with JavaScript is that, because it is a cross-platform scripting language that runs inside a browser, it has very limited programmatic access to computer hardware. As a result, many processing steps are needed to present stimuli and receive user input, making it very difficult to accurately measure RTs. Prior online behavioral research (e.g., Reimers & Stewart, 2014) has used a time-polling subroutine—the *getTime*() method of the *Date* object—that has millisecond resolution but not necessarily millisecond precision,[4] especially on Windows PCs. Although previous work has demonstrated that this subroutine (here called the "date method") is accurate enough to support the replication of some fairly subtle effects, including RT differences between compatible and incompatible trials in the flanker task (Crump et al., 2013), the technique has not yet been used in conjunction with auditory stimuli.

This is perhaps with good reason: The Web development community has historically struggled (Wilson, 2013) with the synchronization of auditory events with other forms of interaction, because of the complexities associated with playing audio on the Web that we mentioned in the introduction. For example, pseudocode for a naive implementation of RT measurement for the ALD task in Experiment 1 might look as follows:

1. Wait 250 ms as an ISI
2. Start playing audio stimulus
3. Record the stimulusStart time with the date method
4. Wait for user response
5. Record the responseTime time with the date method
6. reactionTime = responseTime – stimulusStart

Unfortunately, there is no way to guarantee that the stimulus start time measured by the date method is aligned with the actual onset of the auditory stimulus, because an unknown amount of time lag separates when an audio component is asked to play and when it actually begins to do so (see, e.g., Psychology Software Tools, 2014; Smus, 2012).

However, a widely supported[5] high-level sound interface for JavaScript now exists, the Web Audio API, that can address this problem (Adenot & Wilson, 2015). Among other features, the Web Audio API includes access to the soundcard's clock via the *currentTime* property of the

[4] Here "resolution" refers to the number of digits in the value returned by the function, whereas "precision" is the measurement's tolerance (guaranteed ± neighborhood) relative to the true value. In other words, not all digits provided by the *getTime*() method are necessarily meaningful.

[5] Current versions of Google Chrome, Mozilla Firefox, Safari, and Opera all support the Web Audio API, and support in Internet Explorer is planned. See http://caniuse.com for detailed support information.

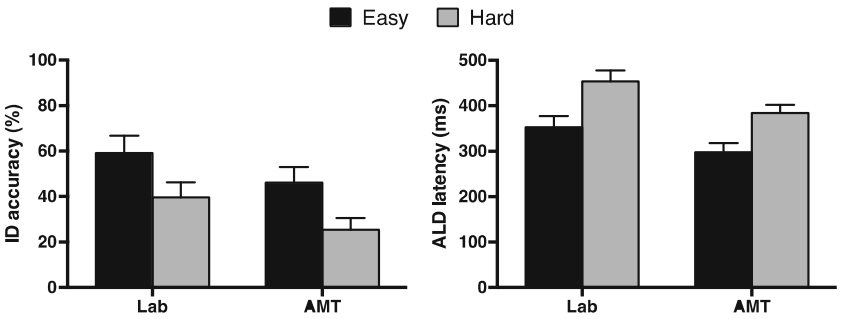

Fig. 5 Influences of lexical difficulty on identification-in-noise (ID) accuracy and Auditory lexical decision (ALD) latency

AudioContext object (referred to here as the "audio method"), as well as dedicated audio scheduling. This feature allows the programmer to plan sounds at specific points in the soundcard's time course. Pseudocode for the actual ALD implementation used in Experiment 1 is as follows:

1. Record the currentTime time with the audio method
2. stimulusStart = currentTime + 250 ms (the ISI)
3. Schedule the next audio stimulus to begin playing at stimulusStart
4. Wait for user response
5. Record the responseTime time with the audio method
6. reactionTime = responseTime – stimulusStart

By implementing measurement in this fashion, it may be possible to gain more accurate RT measurements. It was the purpose of this experiment to compare the accuracy of measurements provided by the audio method to those provided by the date method.

Method

In order to compare the two timing methods, we conducted date- and audio-method versions of the ALD task from Experiment 1 with a closed-loop response system. This enabled us to precisely record the actual RTs and compare them to those measured by the JavaScript implementations. The codebase of the ALD task was kept as close as possible to that of the Experiment 1, so as to ensure that the simulation took place in a realistic computational environment, but some simplifying alterations were necessary in order to guarantee accurate control measurements. The stimuli were replaced with a single, 650-ms-long (approximately the mean length of our stimuli), 440-Hz pure tone to provide an unambiguous stimulus start time. Standard ALD responses (two keys, one for "word" and the other for "nonword") were simplified to a single key.

Two computers were used in this experiment: a "test" machine that ran the modified experiment, and a "control" machine that recorded the time course of the stimuli and responses. Responses were automated by an external Arduino device attached to a double-position, double-throw relay. One of the relay's poles closed the contacts of a key on the test computer's keyboard. The other closed a circuit that generated a small spike in an audio channel of the control computer's line-in soundcard input. The other channel of the control computer's line-in was attached to the headphone jack of the test computer. Because the dual-acting relay simulated a participant's response and generated a small waveform at the same time, by recording the control computer's line-in stereo input, we obtained a time-locked record of stimulus presentation and "participant" responses. The response device provided RTs distributed approximately uniformly between 500 and 1,000 ms. The control computer's line-in input was recorded with Audacity.

Due to the vast number of different computer models currently being used to access the Internet, it is impossible to investigate timing behavior on every device. Instead, many chronometry studies (Reimers & Stewart, 2014; Simcox & Fiez, 2014) approximate this diversity by gathering data from a set of typical hardware configurations, operating systems, and Web browsers. However, because the present study was concerned with the relative difference in performance between the two timing functions of the same software language, we only considered a single hardware and software setup: a Lenovo X220 laptop running Google Chrome on 64-bit Windows 8 with 4 GB of RAM and an Intel Core i5-2520 M CPU clocked at 2.50 GHz.

This machine was chosen because it is in the range that AMT participants might be expected to use,[6] but is likely to render differences between the timing methods that are smaller than will typically be observed. An older computer, a dedicated and/or higher-quality soundcard, or a Web browser with a slower JavaScript engine would likely yield the same or

[6] Although up-to-date operating system statistics for AMT users are generally difficult to find, a first approximation is provided by Experiment 1, in which 84 % of the participants used Windows. This is consistent with the findings of Reimers and Stewart (2014), who reported that 85 %–90 % of the participants in online experiments (including AMT and other platforms) use Windows.

starker differences between the two timing methods. Furthermore, in a timing study concerned with visual stimuli, Reimers and Stewart (2014) found "no obvious systematic effect" of browser type on RT measurements, so browser choice may be largely irrelevant. In summary, any differences between the two methods revealed here would represent a conservative estimate of the difference between the two methods across platforms.

The experiment was run with each timing method twice: first while the processor was under low load (approximately 5 % processor use), and then while it was under high load (approximately 65 % processor use). The high-load condition was included to simulate a participant who was running other software during the experiment. In keeping with Simcox and Fiez (2014), Prime95 was used to generate processor load. A total of 250 trials were conducted in each of the four conditions (low and high processor load for both the audio and date methods).

Results and discussion

For each trial in both timing conditions and at both load levels, the time between the onset of the stimulus and the response was manually measured using Audacity. For each trial, the RTs measured by the control computer (i.e., the actual RTs) were subtracted from those measured by the test computer (the RTs measured by the experimental code), to obtain a measurement error (see Table 5 for descriptive statistics). The test (measured) RTs were longer than the control (actual) RTs on every trial (on average by 59 ms, $SD = 11$ ms). Although this overestimation of RT may seem large, it is within the range of latencies reported by prior research. For instance, Reimers and Stewart (2014) found RT overestimation by 30–100 ms when using JavaScript and Flash timing methods across a range of devices. Plant and Turner (2009) found significant lags in two contributors to this latency: keyboards (delays up to 34 ms) and speaker systems (delays up to 37 ms). Psychology Software Tools Inc., the makers of E-Prime, a leading in-lab stimulus presentation software program, found an even greater range of speaker system lags, up to a mean of 368 ms for some hardware and firmware combinations (Psychology Software Tools, 2014).

Table 5 Means and standard deviations for measurement errors, measured by the date and audio timing methods in two processor load conditions

	5 % load		65 % load	
	Mean	*SD*	Mean	*SD*
Date method	54.67	5.88	61.34	20.21
Audio method	60.85	4.29	60.33	4.09

Times are in milliseconds.

The date method provided measurements closer to the actual values than did the audio method, $F(1, 996) = 13.94$, $MSE < .001$, $p < .001$, $\eta_p^2 = .02$, and measurements were closer to the actual values in the low-load than in the high-load condition, $F(1, 996) = 19.72$, $p < .001$, $MSE < .001$, $\eta_p^2 = .02$. In addition, a significant Method × Load interaction emerged, $F(1, 996) = 27.00$, $MSE < .001$, $p < .001$, $\eta_p^2 = .03$. Planned comparisons indicated that the cause for the interaction was a significant effect of load for the date method, $t(498) = -5.00$, $p < .001$, Cohen's $d = -0.45$, but not for the audio method, $t(498) = 1.39$, $p = .17$, Cohen's $d = 0.12$. These analyses demonstrate that the effect of load was greater for the date than for the audio method. Moreover, Levene's test for equality of variances showed that the variance in measurement error for the audio method was significantly smaller than those related to the date method for both the low-load, $F(1, 498) = 26.89$, $p < .001$, and high-load $F(1, 498) = 4.74$, $p = .03$, conditions.

How should we compare these timing methods? The two most salient criteria are the mean and variance of each method's timing error, but the first criterion is rarely relevant: Actual RTs collected from uncontrolled timing systems should not be used alone to support theoretical results, because the amounts of lag in these systems are inconsistent across participants, and thus cannot be accounted for (unlike in carefully controlled laboratory settings). Instead, RT measurements in this context are used comparatively; that is, the result sought is the difference between RTs in two separate conditions (e.g., lexically hard words vs. easy words) on a participant-by-participant basis. When RTs from the same participant (and, consequently, the same computer system) are treated in this way, the error value (latency) is largely removed via subtraction. For within-subjects studies evaluating item differences (e.g., evaluating the influence of lexical variables on word recognition), differences in measurement error due to hardware or load across participants will affect all words equally, and therefore will not systematically bias the results. Using statistical techniques such as mixed-effect models that include random effects for participants can also account for participant variability in measurement errors.

Therefore, for most analyses, variance in errors is the critical statistic for comparing measurement methods. Examining Table 5, we see that in comparison to the date method, the audio method results in more robust measurements, due to its lower variance. This is especially true when systems are under computational stress: Note that the audio method's *SD* appears to be minimally affected by an increase in processor load. Also reassuring is the fact that these *SD*s are close to the range produced by popular in-lab experiment software packages when used with a computer keyboard (rather than a specialized response device). For example, Schubert, Murteira, Collins, and Lopes (2013) found that E-Prime, DMDX, Inquisit, and Superlab have respective measurement error *SD*s of 3.30, 3.18, 3.20, and 4.17 ms. It is important to note,

however, that the values listed in Table 5 are not universal results; they are specific to the hardware, firmware, and software combination that was used. Instead, these results represent a general pattern of increased measurement quality provided by the audio method relative to the date method. Due to our conservative choice of technology, this difference is expected to remain the same or become more pronounced under other circumstances.

For the experimenter planning an online study requiring measurements of reaction speeds to auditory stimuli, these results should be reassuring. Under light processor load, both methods are roughly equivalent and not far from the fidelity provided by in-lab setups. The decision of which method to use depends primarily on whether it is more important to support Internet Explorer or to measure RTs in a way that is resilient to varying processor loads. Although current browser statistics for AMT users are difficult to find, Internet Explorer's relatively small market share on the Internet as a whole (13 % including or 19 % excluding mobile devices; StatCounter, 2015), combined with anecdotal evidence that AMT users prefer other Web browsers, suggests that sacrificing universal browser support may not significantly affect the results. This, along with the fact that the Web Audio API is designed specifically for situations like these, motivates the authors' belief that the advantages of the audio method in general outweigh those of the date method.

General discussion

These findings demonstrate strong consistencies in the relative accuracies and latencies of spoken word data collected in the lab and online. In addition, the results show that lab and online data are very similarly correlated with well-established lexical variables. For researchers concerned with modeling spoken word recognition or whose primary focus is evaluating stimulus-level differences, these results suggest that AMT can be an effective venue for data collection. In addition to the fact that these data may be obtained more cheaply and quickly, data collected online may have other distinct advantages for research on spoken word recognition.

As we described in the introduction, the lack of environmental control inherent in online research is often interpreted as a limitation. Yet, in the context of spoken word recognition research, it may actually be advantageous. For example, if researchers are seeking to evaluate the confusability of word pairs or the intelligibility of speech tokens, having diverse listening situations will yield a better approximation of general confusability and intelligibility than do data obtained from stimuli presented in a carefully controlled setting. Therefore, the conclusions drawn from online experimentation may be expected to be more robust and generalizable to natural settings than lab-collected findings.

In addition to environmental variability, participant variability may be valuable for research on spoken word recognition. The growing body of literature that demonstrates the relationship between cognitive abilities and language-processing ability (Benichov, Cox, Tun, & Wingfield, 2012) suggests that college students, a population that does not represent the general population cognitively, may not be expected to represent the general population in language-processing abilities. Furthermore, AMT provides the opportunity to attract participants with a broader range of linguistic backgrounds and experiences, providing a richer participant source for research concerned with accented speech or cross-cultural language processing.

The results of Experiment 2 demonstrate that the Web Audio API recently adopted by popular browsers can indeed provide more accurate time measurements. Although some delay is unavoidable, the audio time-polling method was found to provide significantly more consistent measurements, especially in the high-processor-load condition. The data support the use of the Web Audio API's timing and audio scheduling by researchers hoping to investigate potentially subtle effects related to auditory perception.

Although online research is promising in many regards, it is probably not yet well-suited for certain auditory perception tasks. Given the current technology, it appears that collecting measurements near hearing thresholds or presenting stimuli that require precise control over auditory amplitude will be difficult. However, Cooke et al. (2013) proposed a possible solution, by asking additional questions of participants such as the level of noise that they completed the task in and whether they were listening through headphones or speakers. Future studies could consider filtering participants on the basis of their responses to these types of questions. Another approach could be to present participants with a pretest in which they completed a two-alternative forced choice task for the detection of stimuli at varying intensities. This could give a direct measurement of stimulus detectability, which could be used to approximate the combined influences of the hearing level of the user and the environment in which the study was conducted.

In addition, more research will be required to determine whether and under which circumstances RT experiments concerned with individual differences can be conducted online. In cases in which computer performance is distributed uniformly across groups of individuals, researchers may be able to avoid bias. However, in cases in which this cannot be guaranteed, one must be very cautious. For instance, an online study comparing the RTs of different age groups may yield biased results as a product of computer age (and, therefore, computer performance) being correlated with participant age.

More generally, behavioral studies that are focused on evaluating absolute performance levels would require care when drawing direct comparisons between in-lab and AMT data.

Given the measurement lag that is unavoidable in consumer devices, as well as the work suggesting that performance is likely to be less accurate on AMT than in laboratory measures (Cooke et al., 2013), it is important to acknowledge differences in motivation, environment, technology, and demographics when presenting such data. It is also important to keep in mind that the results of this study do not reflect the additional challenges associated with experiments that employ multimedia stimuli. Tightly synchronizing visual and auditory presentation is difficult, let alone measuring RTs relative to such stimuli. Although the Web Audio API is designed to aid in such circumstances, work beyond the present study will be required to verify the interface's ability to do so in the context of psychological research.

Author note We are grateful to Violet Brown and Emily Massell for assistance with data collection, and to Dennis Barbour, Sarah Meerts, and Julie Neiworth for helpful comments on a previous draft. This research was supported in part by a grant from Howard Hughes Medical Institute to the Carleton College Interdisciplinary Science and Math Initiative (Grant No. 52006286). Portions of this work were presented at the Auditory Perception, Cognition, & Action Meeting in November 2014.

References

Adenot, P., & Wilson, C. (2015). Web audio API (W3C editor's draft). Retrieved January 8, 2015, from http://webaudio.github.io/web-audio-api/

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The english lexicon project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Benichov, J., Cox, L., Tun, P., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability and cognitive function. *Ear and Hearing, 32,* 250–256. doi:10.1097/AUD.0b013e31822f680f.Word

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Buhrmester, M., 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. doi:10.1177/1745691610393980

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1084–1096. doi:10.1037/0278-7393.16.6.1084

Cooke, M., Barker, J., Garcia Lecumberri, M., & Wasilewski, K. (2011). Crowdsourcing for word recognition in noise. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of Interspeech 2011* (pp. 3049–3052). Grenoble, France: International Speech Communication Association.

Cooke, M., Barker, J., Lecumberri, M. L. G., & Wasilewski, K. (2013). Crowdsourcing for speech perception. In M. Eskenazi, G. Levow, & H. Meng (Eds.), *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment* (pp. 137–172). New York, NY: Wiley.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*, e57410. doi:10.1371/journal.pone.0057410

Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 249–283). San Diego, CA: Academic Press.

Field, A. (2009). *Discovering statistics using SPSS*. London, UK: Sage.

Goh, W. D., Suárez, L., Yap, M. J., & Tan, S. H. (2009). Distributional analyses in auditory lexical decision: Neighborhood density and word-frequency effects. *Psychonomic Bulletin & Review, 16,* 882–887. doi:10.3758/PBR.16.5.882

Goldinger, S. D. (1996). Auditory lexical decision. *Language & Cognitive Processes, 11,* 559–568. doi:10.1080/016909696386944

Google, Inc. (2014). reCAPTCHA. Retrieved from www.google.com/recaptcha

Kaiser, A. R. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46,* 390–404. doi:10.1044/1092-4388(2003/032)

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44,* 978–990. doi:10.3758/s13428-012-0210-4

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception, 62,* 615–625. doi:10.3758/BF03212113

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36. doi:10.1097/00003446-199802000-00001

Luce, P. A., Pisoni, D., & Goldinger, S. (1990). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive models of speech processing* (pp. 122–147). Cambridge, MA: MIT Press.

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23. doi:10.3758/s13428-011-0124-6

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. In *Proceedings of Interspeech 2012* (pp. 1708–1711). Grenoble, France: International Speech Communication Association.

McClelland, J. L., Elman, J. L., & Diego, S. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86. doi:10.1016/0010-0285(86)90015-0

McDonnell, J. V., Martin, J. B., Markant, D. B., Coenen, A., Rich, A. S., & Gureckis, T. M. (2012). psiTurk (Version 1.02) [Software]. New York, NY: New York University. Retrieved from https://github.com/NYUCCL/psiTurk

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60,* 355–376.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

Pisoni, D. B. (1996). Word identification in noise. *Language & Cognitive Processes, 11,* 681–688. doi:10.1080/016909696387097

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods, 41,* 598–614. doi:10.3758/BRM.41.3.598

Psychology Software Tools, Inc. (2014). Sound startup latency tests [Software]. Retrieved from www.pstnet.com/eprimestartup.cfm

Reimers, S., & Stewart, N. (2014). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*. doi:10.3758/s13428-014-0471-1

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America, 35,* 200.

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A software library for collecting response latencies in

online studies of cognition. *PLoS ONE, 8*, e67769. doi:10.1371/journal.pone.0067769

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods, 46,* 95–111. doi:10.3758/s13428-013-0345-y

Smus, B. (2012). Developing game audio with the Web audio API. Retrieved from http://html5rocks.com

Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging, 11,* 333–341. doi:10.1037/0882-7974.11.2.333

Sommers, M. S. (2002). Washington University Speech and Hearing Lab Neighborhood Database. Retrieved from http://neighborhoodsearch.wustl.edu/Home.asp

Sommers, M. S., & Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging, 14,* 458–472.

StatCounter. (2015). StatCounter global stats. Retrieved from http://gs.statcounter.com

Strand, J. F. (2014). Phi-Square Lexical Competition Database (Phi-Lex): An online tool for quantifying auditory and visual lexical competition. *Behavior Research Methods, 46,* 148–158. doi:10.3758/s13428-013-0356-8

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *Journal of the Acoustical Society of America, 130,* 1663. doi:10.1121/1.3613930

Turner, J. E., Valentine, T., & Ellis, A. W. (1998). Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision. *Memory & Cognition, 26,* 1282–1291. doi:10.3758/BF03201200

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9,* 325–329. doi:10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers, 36,* 481–487. doi:10.3758/BF03195594

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68,* 306–311. doi:10.1006/brln.1999.2116

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 3,* 387–401. doi:10.1002/wcs.1178

Wilson, C. (2013). A tale of two clocks—Scheduling Web audio with precision. Retrieved from www.html5rocks.com/en/tutorials/audio/scheduling/

Wolters, M. K., Isaac, K. B., & Renals, S. (2010). Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In Y. Sagisaka & K. Tokuda (Eds.), *Proceedings of the 7th Speech Synthesis Workshop (SSW7)* (pp. 136–141). Kyoto, Japan: National Institute of Information and Communications Technology. Retrieved from http://isw3.naist.jp/~tomoki/ssw7/www/doc/ssw7_proceedings_rev.pdf