The Feasibility of a Japanese Crowdsourcing Service for Experimental Research in Psychology

Yoshimasa Majima

Abstract
Recent studies have empirically validated the data obtained from Amazon’s Mechanical Turk. Amazon’s Mechanical Turk workers behaved similarly not only in simple surveys but also in tasks used in cognitive behavioral experiments that employ multiple trials and require continuous attention to the task. The present study aimed to extend these findings to data from Japanese crowdsourcing pool in which participants have different ethnic backgrounds from Amazon’s Mechanical Turk workers. In five cognitive experiments, such as the Stroop and Flanker experiments, the reaction times and error rates of Japanese crowdsourcing workers and those of university students were compared and contrasted. The results were consistent with those of previous studies, although the students responded more quickly and poorly than the workers. These findings suggested that the Japanese crowdsourcing sample is another eligible participant pool in behavioral research; however, further investigations are needed to address issues of qualitative differences between student and worker samples.

Keywords
non-Caucasian crowdsourcing pool, online experiment, reaction time measurement, Qualtrics, experimental cognitive research

Researchers in the behavioral and social sciences (such as psychology, linguistics, economics, and political science) have begun to collect data from online surveys and online experiments using participants recruited from an online labor market, a procedure known as crowdsourcing. Recruiting participants from crowdsourcing services has certain advantages over using traditional samples consisting of university students. In particular, utilizing crowdsourcing as a participant pool enables researchers to access “large” amounts of data from a “diverse” sample in a “short period” of time at a relatively “low cost” (Mason & Suri, 2012). Although there are many extant crowdsourcing services, Amazon’s Mechanical Turk, often abbreviated as MTurk or AMT, is the most popular system among behavioral researchers.

AMT
AMT is designed to provide a platform that helps employers (requesters) hire employees (workers) and compensate them in exchange for completing work known as human intelligence tasks (HITs). As a participant pool for academic research, this platform provides researchers with a sizable number of potential participants from diverse backgrounds and offers a one-stop integrated system that includes posting HITs, recruiting participants, and processing payments. In addition, using AMT is typically less expensive than standard laboratory experiments or other online recruitment methods. Furthermore, AMT participants are completely anonymous to researchers, and Internet-based experiments can thus minimize so-called experimenter effects. Finally, the codes for online surveys and experiments can be shared easily with other researchers who wish to replicate studies using a different sample (Crump, McDonnell, & Gureckis, 2013).

Recently, some studies have examined the usefulness of AMT as a data collection tool in research in psychology and the behavioral and social sciences. In a typical examination, the demographic status and various psychological properties of workers were contrasted with those of students or age-matched community samples. These studies indicated that most AMT workers are residents of the United States—the next largest place of residence is India—and that workers were different from non-Internet participants in terms of age, education level, employment status, and religious and political attitudes (for recent reviews, see Chandler & Shapiro, 2013).

Corresponding Author:
Yoshimasa Majima, Department of Psychology for Well-Being, Hokusei Gakuen University, 2-3-1, Oyachi Nishi, Atsubetsu-ku, Sapporo 0048631, Japan.
Email: majima.y@hokusei.ac.jp

1Hokusei Gakuen University, Sapporo, Japan

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://www.creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
In addition, AMT workers were less extraverted or emotionally stable and had lower self-esteem than the student and community samples (e.g., Behrend, Sharek, Meade, & Wiebe, 2011; Goodman, Cryder, & Cheema, 2013). The recent comparison of a large AMT sample with online and face-to-face benchmark samples also exhibited that AMT workers were lower in extraversion than two contrast groups and higher in openness compared to the online benchmark sample (Clifford, Jewell, & Waggoner, 2015); however, they were not different in terms of other Big Five traits.

Regarding attention to study materials, Goodman et al. (2013) showed that AMT workers were more likely to miss important instructions than students. However, failures in attentional checks were mainly found in English as a Second Language (ESL) and non-U.S. participants; hence, inattentive responses are partly due to the lack of language proficiency. Furthermore, Hauser and Schwartz (2016) revealed that AMT workers demonstrated better performance than students on the instructional manipulation checks. Therefore, it is reasonable to assume that the AMT workers comprise a more attentive sample than (at least same as) the student sample. In addition, Paolacci, Chandler, and Ipeirotis (2010) noted that the workers and students were not different with regard to self-reporting numeracy and attention to survey instructions. They also reported that both participants exhibited a significant judgment bias, such as the framing effect, conjunction fallacy, and outcome bias. Similar results were reported by Goodman et al. (2013), in which the workers and students showed similar classic decision-making biases, such as risk preference and the certainty effect.

These validation studies indicated that AMT workers are slightly different in demographic status and certain personality traits compared with the traditional subject pool; however, the workers were sufficiently attentive to the task instructions, and the participants from the two samples solved judgment and decision-making tasks in a similar way. Therefore, studies suggest that using AMT with the online survey platform is a useful alternative to traditional paper-and-pencil questionnaire surveys.

Crump et al. (2013) stimulated the validation studies of AMT in online cognitive behavioral experiments, such as the Stroop, Flanker, Simon, attentional blink, and category learning experiments. These experimental tasks, which are generally adopted in laboratory studies, include multiple trial design, accurate control of stimulus presentation, and reaction time (RT) measurement. Crump et al. (2013) aimed to qualitatively replicate significant findings from laboratory studies in online experiments using AMT participants. Their study did not make a direct comparison of the performance of lab and online participants; however, the results were in line with those obtained from laboratory studies and suggested that the online crowdsourcing pools are also eligible for cognitive behavior studies.

CrowdWorks as a Japanese Participant Pool for Behavioral Studies

Previous validation studies have shown that data obtained from AMT workers were mostly reliable, and having access to a large participant pool—often in a matter of hours—has encouraged researchers to use AMT as a platform for collecting behavioral data. However, despite its advantages, using AMT has certain limitations. First, AMT workers are more diverse than university students, but demographic surveys have repeatedly shown that AMT workers are predominantly Caucasian residents in the United States, followed by Indian workers (e.g., Behrend et al., 2011; Goodman et al., 2013; Paolacci et al., 2010). In other crowdsourcing services, such as Prolific (https://prolific.ac/) or Clickworker (https://www.clickworker.com/), most potential participants are also Caucasian residents of the United States and European countries. Consequently, if researchers wish to extend findings to samples composed of other ethnicities or nationalities in their online studies, they must gain access to a different participant pool. The second issue involves AMT workers’ agreements to their payment service. Until now, requesters were required to use the Automated Clearing House (ACH) system and a valid ACH-enabled bank account in the United States. This requirement also hinders researchers outside the United States from using AMT as a participant pool.

Of course, several studies have explored the viability of other crowdsourcing pools, such as Clickworker and CrowdFlower (e.g., Lutz, 2016; Peer, Samat, Brandimarte, & Acquisti, 2015). Although the results indicated mixed results, certain non-AMT crowdsourcing pools may be practical alternatives to AMT. However, it is still unclear whether collecting participants from other crowdsourcing samples, particularly from non-Caucasian samples, is also a viable option in behavioral studies. The viability of non-Caucasian samples must be investigated to facilitate the use of crowdsourcing as a participant pool for research in psychology and other behavioral and social sciences. Recently, Majima, Nishiyama, Nishihara, and Hata (2017) compared workers from a Japanese crowdsourcing service (CrowdWorks; described later in detail) and university students in terms of their demographic data, personality traits, attention to instructions, and reasoning skills. Majima et al. (2017) found that workers and students were different in terms of demographic status and certain personality traits; however, they demonstrated similar performance on various reasoning tasks. For example, CrowdWorks workers were older, had more extensive work experience, and were less extraverted and agreeable but were more conscientious than students. In addition, although workers were more attentive to instruction than students, they responded similarly to other reasoning tasks. The results were generally compatible not only with the existing AMT validation studies but also with other offline studies using Japanese participants.
The present study aimed to extend the findings from a previous validation study of AMT as a participant pool for online cognitive behavioral experiments to a Japanese (i.e., non-Caucasian) sample collected from a Japanese crowdsourcing service. Specifically, we focused on five classical cognitive experiments adopted in Crump et al. (2013)’s study: the Stroop effect (MacLeod, 1991; Stroop, 1935), the Flanker effect (Eriksen & Eriksen, 1974), the task-switching cost (Kiesel et al., 2010; Monsell, 2003), the Simon effect (Simon, Sly, & Vilapakkam, 1981), and visual-cueing experiments (Posner & Cohen, 1984). The five tasks were chosen because they were frequently used in the field and hence were suitable for empirical validation. Furthermore, the laboratory studies with the Japanese sample found similar results as with Caucasian samples (e.g., the Stroop effect, Fukushima, Miki, Yokouchi, & Hiroyasu, 2013; Yamazaki, 1985; the Flanker effect, H. Tanaka, Masaki, Takasawa, & Yamazaki, 2002; task-switching cost, Harada, Asano, Suto, & Hasher, 2010; Umebayashi & Okita, 2011; the Simon effect, Watanabe & Yoshizaki, 2016; visual-cueing task, Matsuda, 2012). In addition, this study also compared data from crowdsourcing participants and those from the traditional subject pool (i.e., university students) using the same online experimental platform. In particular, this study compared workers to students in terms of their response speed and accuracy in these tasks.

The present study drew student participants from two middle-sized Japanese universities and crowdsourcing participants from the Japanese crowdsourcing service CrowdWorks (http://crowdworks.jp). We adopted CrowdWorks as the present participant pool because of several similarities to AMT. For example, CrowdWorks has a sufficiently large pool of registered workers (approximately 700,000 workers as of late 2015, when the experiments had been conducted) for empirical studies. Second, the predominant workers in CrowdWorks are native Japanese speakers; hence, it offers a participant pool with different ethnic groups than AMT. Third, CrowdWorks has a similar one-stop integrated system as AMT, and it charged no commission fee for micro-tasks.

**General Method of the Experiments**

All tasks were implemented using the online survey software administered by Qualtrics, together with the QRTEngine (Qualtrics Reaction Time Engine; Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015), which enables a response latency measurement in a Qualtrics environment. All participants who were drawn from two (crowdsourcing and student) samples completed the experiments in the same online environments.

In each experiment, approximately 100 crowdsourcing and 100 student participants were recruited. The crowdsourcing participants were recruited from the Japanese crowdsourcing service CrowdWorks. In the present study, no exclusion criterion, such as demographic requirements and acceptance ratio of completed tasks, was applied when recruiting participants. However, the demographic results showed that almost all were native Japanese speakers, except for only one worker (Chinese speaker) who participated in Experiment 5. The participants received 50 JPY (approximately $.45) for their participation in each experiment. The student participants were recruited from two different psychology classes in two middle-sized universities in Japan. They received extra credit for their participation. The data from some participants (the numbers were different between the tasks) were excluded from the analysis because these participants did not complete the task or failures occurred in the data recording (for additional details, see Table S1 in the supplemental material). We enabled the Qualtrics restriction feature to prohibit both crowdsourcing workers and students from participating in a single task more than once. However, we did not track the participants’ identities; therefore, there may be overlap in participants across the five experiments.

We posted a link to a survey page administered by Qualtrics to the CrowdWorks tasks. The student participants received a leaflet with a link to the survey page. Unless otherwise indicated, each task consisted of approximately 100 trials and lasted approximately 5 min.

**Experiment 1: Stroop Effect**

Experiment 1 administered a classic cognitive experiment, the Stroop interference experiment. In a typical Stroop task, the participants are asked to identify the color in which a color word is displayed (MacLeod, 1991; Stroop, 1935). Across the variants of the response mode in identification, congruent (e.g., “red” written in a red font) word-color pairs result in faster response times and lower error rates compared with incongruent (“red” written in a green font) word-color pairs.

**Method**

*Stimuli and design.* Experiment 1 followed the same stimuli, design, and procedure as that in Barnhoorn et al. (2015) and Crump et al. (2013), except that the instructions and word stimuli were written in Japanese. The stimuli consisted of red, green, blue, or yellow colors paired with the respective Japanese words. Four congruent and 12 incongruent word-color pairs were used, resulting in a total of 96 trials with 48 congruent and 48 incongruent pairs. The words were displayed in 50-point font at the center of a webpage with a white background.

*Participants and procedure.* One hundred students (42 female, 90% right-handed, mean age = 19.8 years) and 99 crowdsourcing workers (51 female, 94% right-handed, mean age = 37.7 years) participated in this experiment. After the participants read the general instructions, they were asked to
Results and Discussion

Four students and three CrowdWorks workers were excluded because their response accuracy levels were below 80%. Only the RTs for the correct trials were submitted for the following analyses. Furthermore, after the RTs of each response were submitted to an outlier analysis (Selst & Jolicoeur, 1994), 1.2% of the observations were excluded (on average, 0.7 trials per student participant and 1.4 trials per worker participant were excluded; for details, see Table S2 of the supplementary materials). The mean RTs and error rates for each participant were submitted to separate two-way mixed-design analyses of variance (ANOVA) with the samples (UNIV = students vs. CW = crowdsourcing participants) as the between-subject variable and congruency as the within-subject variable. Figure 1 shows the mean RTs and error rates as a function of sample and congruency.

In both samples, the RTs were faster for congruent items (UNIV = 688 ms vs. CW = 873 ms, mean across samples = 781 ms, standard deviations [SDs] = 146.3, 183.8, 190.0, respectively) than incongruent items (UNIV = 771 ms vs. CW = 970 ms, mean across samples = 870 ms, SDs = 168.3, 217.2, 218.0, respectively). F(1, 190) = 270.2, p < .001, η² = .22. The error rates were low overall; however, we found a significant main effect of the sample (UNIV = .042 vs. CW = .023, SDs = 0.041, 0.039), F(1, 190) = 16.2, p < .001, η² = .08, and a significant main effect of congruency (congruent vs. incongruent .042, SDs = 0.033, 0.047), F(1, 190) = 40.2, p < .001, η² = .17. The sample-congruency interaction was not significant (F < 1).

The results of the RT measurements were consistent with those of previous studies (Barnhoorn et al., 2015; Crump et al., 2013; Fukuhara et al., 2013). The RTs for congruent items were faster than those for incongruent items. In addition, the present participants responded faster compared with the AMT workers in Crump et al.’s (2013) study (Table S4 of supplementary materials exhibited mean RTs of the present and the previous studies). It was also found that student participants were faster than both the present and previous crowdsourcing samples but produced relatively higher error rates. However, the RTs of the student participants were slower compared with those of the Japanese laboratory participants (e.g., Fukuhara et al., 2013; for additional details, see Table S4).

Experiment 2: Flanker Effect

The Flanker effect is another form of cognitive conflict that requires dissociating relevant information from irrelevant contextual information (Eriksen & Eriksen, 1974). In a typical Flanker task, the participants are presented with a series of strings, such as “ffff” or “ffhh,” and are asked to identify the letter at the very center of the string as quickly and accurately as possible. The strings are either compatible or incompatible. The compatible strings (e.g., “ffff”) consist of a series of identical letters; the incompatible strings (e.g., “ffhh”) consist of a series of identical contextual letters and a target letter that differs from the contextual letters. Because the incompatible contextual letters hinder the processing of the target, the RTs are generally slower for incompatible trials.
Method

Stimuli and design. Experiment 2 also followed the same stimuli, design, and procedure as that done in Crump et al. (2013). The Flanker items were strings that consisted of the lowercase letters “f” and “h.” The compatible strings were either “fffff” or “hhhhh.” The incompatible strings were either “ffhff” or “hhfhh.” The items were presented in black 50-point font at the center of a white page. The participants were asked to identify the target letters by pressing either the F or H key.

A total of 100 trials were run, with 50 compatible and 50 incompatible items. In each trial, one of the four items was presented at random, and the presentation order was different for each participant.

Participants and procedure. One hundred students (44 female, 88% right-handed, mean age = 19.8 years) and 99 workers (49 female, 95% right-handed, mean age = 38.0 years) participated in the experiment. The general procedure of the experiment was identical to that of Experiment 1. Each trial began with a fixation cross displayed for 1,000 ms, followed by a 500 ms blank screen. The Flanker stimulus was then displayed and remained onscreen until the participants pressed either the F or H key. The participants were asked to press one of the two keys as quickly and accurately as possible. Finally, accuracy feedback was presented for 500 ms.

Results and Discussion

All participants responded correctly to more than 80% of the trials; hence, none of the participants was excluded from the analyses. The same outlier analysis as Experiment 1 excluded 0.9% of the observations (UNIV = 0.4 vs. CW = 1.2 trials per participant).

The mean RTs and error rates for each participant were submitted to a similar Sample × Compatibility mixed-design ANOVA. Figure 2 shows the mean RTs and error rates as a function of sample and compatibility. The participants responded faster for compatible items (UNIV = 531 ms vs. CW = 624 ms, mean across samples = 577 ms, SDs = 94.5, 147.0, 131.8, respectively) than for incompatible items (UNIV = 568 ms vs. CW = 669 ms, mean across samples = 619 ms, SDs = 95.1, 128.7, 123.8, respectively), F(1, 197) = 136.5, p < .001, ηp2 = .41. Furthermore, UNIV was faster than CW, F(1, 197) = 35.4, p < .001, ηp2 = .15. The Sample × Compatibility interaction was not significant, F(1, 197) = 1.3, p = .25.

The error rates were low overall, as in the Stroop test; however, we also found a significant main effect of the sample (UNIV = .040 vs. CW = .022, SDs = 0.044, 0.030), F(1, 197) = 16.8, p < .001, ηp2 = .08, and a main effect of compatibility (compatible = .022 vs. incompatible = .040, SDs = 0.030, 0.044), F(1, 197) = 41.8, p < .001, ηp2 = .18. The sample-congruency interaction was not significant, F(1, 197) = 1.9, p = .17.

The pattern of the results was quite similar to that of Experiment 1 and the previous study. The present results replicated the significant Flanker effect, that is, a higher error rate and slower RTs for incompatible items. Furthermore, the student sample showed relatively higher error rates and faster RTs than the CW participants as shown in the Stroop task. Once again, the RTs of the present participants were slightly faster than those of the AMT workers but were slower than those of the Japanese laboratory participants (H. Tanaka et al., 2002; see Table S4).

Experiment 3: Task Switching

The task-switching paradigm has been used to explore cognitive control mechanisms (Kiesel et al., 2010). In the task-switching experiment, the same task repetition generally results in faster and more accurate responses, and task alteration impedes responses. This asymmetry is often referred as a “switch cost.” The present experiment aimed to replicate the switch cost effect using the standard experimental procedure. The stimulus set, designs, and task procedure were same as those used in Crump et al. (2013). In each trial, the participants were given one of two task cues and a single target digit. The target digits were integer numbers ranging from 1 to 9, excluding 5. The task cues were either “EVEN/ODD” or “LARGER/SMAller THAN 5,” written in Japanese. The participants were asked to judge the target digit in terms of the presented cue and press the designated button as quickly and accurately as possible. Two tasks were randomly presented to the participants. Whether the task was repeated or switched was decided based on the previous trial.
**Method**

*Stimuli and design.* The target items were one of the following integer numbers: 1 to 4 or 6 to 9. The target and task cue were presented in black 50-point (target) or 20-point (task cue) font on a white background. The “EVEN/ODD” task was completed by pressing the A or S key, and the “LARGER/SMALLER” task was completed by pressing the L or K key, respectively. Accuracy feedback was presented in black 30-point font. A total of 96 trials consisted of 48 even/odd and 48 larger/smaller task cues. The task cues and target digit were randomly assigned to a given trial.

*Participants and procedure.* One hundred students (47 female, 92% right-handed, mean age = 19.8 years) and 100 workers (47 female, 94% right-handed, mean age = 38.7 years) participated in the experiment. The general procedure used in Experiments 1 and 2 was also employed. Following a fixation cross displayed for 1,000 ms and a 500 ms blank screen, the target digit and task cue remained on the screen until the participants pressed one of the four keys. The participants were asked to respond as quickly and accurately as possible. Finally, the accuracy feedback was presented for 500 ms.

**Results and Discussion**

Four student participants were excluded because of the low accuracy of responses. Furthermore, the same outlier analysis as Experiments 1 and 2 for the RTs also excluded 1.3% of the observations (UNIV = 0.6 vs. CW = 1.7 trials per participant).

The mean RTs and error rates for each participant were submitted to separate Sample × Switch mixed-design ANOVAs. Figure 3 shows the mean RTs and error rates as a function of the sample and trial type. The participants were generally faster in the repeated trials (UNIV = 970 ms vs. CW = 1,224 ms, mean across samples = 1,099 ms, SDs = 220.2, 355.6, 322.5, respectively) than in the switched trials (UNIV = 1,042 ms vs. CW = 1,306 ms, mean across samples = 1,174 ms, SDs = 207.4, 334.9, 308.9, respectively), F(1, 194) = 88.0, p < .001, $\eta_p^2 = .31$. Similar to Experiments 1 and 2, UNIV was faster than CW, F(1, 194) = 41.1, p < .001, $\eta_p^2 = .18$. The Sample × Switch interaction was not significant, F < 1.

Again, the error rates were low overall; however, significant main effects of the sample (UNIV = .079 vs. CW = .034, SDs = 0.059, 0.039) and switch (repeated = .045 vs. switched = .067, SDs = 0.047, 0.060) were found, Fs(1, 194) = 60.5, 35.6, ps < .001, $\eta_p^2$s = .24, .16, respectively. Furthermore, the sample-switch interaction approached significance, F(1, 194) = 3.5, p = .062, $\eta_p^2 = .02$. The switch-repeat difference was somewhat higher in the students, F(1, 194) = 30.2, p < .001, $\eta_p^2 = .14$, than in the workers, F(1, 194) = 8.6, p = .004, $\eta_p^2 = .04$.

To summarize, the task-switching cost was replicated in the present experiment. In addition, the present results also showed that the students responded faster but less accurately than the CW participants, as in the Stroop and Flanker experiments. Although these students responded faster than the CW workers and AMT workers, their RTs were slightly slower than those of previous Japanese students who participated in another sort of laboratory experiment on task-switching cost (Harada et al., 2010; see Table S4).

**Experiment 4: Simon Effect**

The Simon effect is a phenomenon in which responses are faster and more accurate when the stimulus appears in the same relative position or direction as a response (Proctor & Lu, 1999; Simon et al., 1981). In a typical Simon task, visual targets are presented in one of two spatial locations, and the participants are asked to respond by pressing the designated buttons. Generally, RTs are faster when the visual target is located in the same direction as the correct response. For example, when participants are asked to press the key on the left side when the visual target (e.g., “green” circle) appears on screen, they respond faster when the target is located on the left side of the screen than on the right side. A similar procedure and stimulus set as that followed in Crump et al. (2013) was applied to this task.

**Method**

*Stimuli and design.* The visual targets were 100-px green and red squares. The participants were asked to press S for the red square and K for the green square. The display consisted of three 150-px white-filled squares with black borders as
placeholders and a white background. Each placeholder was separated by 170 px. The targets appeared in either the right or left placeholder. The fixation cross always appeared in the center.

A total of 100 trials, consisting of 50 spatially compatible trials (25 red squares appeared on the left side and 25 green square appeared on the right side) and 50 spatially incompatible trials (25 red squares on the right and 25 green squares on the left), were performed by each participant. For each trial, the target color and location were assigned randomly.

**Participants and procedure.** The participants consisted of 100 students (46 female, 94% right-handed, mean age = 19.9 years) and 100 workers (53 female, 94% right-handed, mean age = 37.5 years). For each trial, following the 1,000 ms fixation cross and 500 ms blank, the target was displayed either in the left or right side placeholder until the participants pressed the key. When participants pressed the key, the target and placeholders were removed immediately, followed by an accuracy feedback display for 500 ms.

**Results and Discussion**

One student participant was excluded because of low response accuracy (correct ratio < .80). Furthermore, the outlier analysis excluded 1.1% of observations (UNIV = 1.0 vs. CW = 1.1 trials per participant).

The mean RTs and error rates for each participant were submitted to separate Sample × Spatial compatibility mixed-design ANOVAs. Figure 4 shows the mean RTs and error rates as a function of sample and compatibility. The RTs were faster for compatible targets (UNIV = 472 ms vs. CW = 538 ms, mean across samples = 505 ms, SDs = 76.6, 92.8, 91.2, respectively) than for incompatible targets (UNIV = 487 ms vs. CW = 567 ms, mean across samples = 527 ms, 71.2, 93.6, 92.2, respectively), $F(1, 197) = 67.6, p < .001, \eta^2_p = .26$. Similar to Experiments 1 to 3, UNIV was faster than CW, $F(1, 197) = 39.6, p < .001, \eta^2_p = .17$. The Sample × Compatibility interaction was also significant, $F(1, 197) = 6.7, p = .01, \eta^2_p = .03$. The analyses of the simple main effects indicated that the compatible-incompatible difference was somewhat higher for CW participants than for the students, $F$s(1, 197) = 58.8, 15.7, $p$s < .001, $\eta^2$s = .23, .07, respectively.

An analysis of the error rate showed a significant main effect of compatibility (compatible = .028 vs. incompatible = .041, SDs = 0.037, 0.046), $F(1, 197) = 14.7, p < .001, \eta^2_p = .07$, and sample (UNIV = .046 vs. CW = .023, SDs = 0.046, 0.035), $F(1, 197) = 26.8, p < .001, \eta^2_p = .12$. The Sample × Compatibility interaction was not significant, $F < 1$.

The Simon effect was also replicated in the present study, and the participants’ responses were similar to those from previous studies. Crump et al. (2013) mentioned that the RTs of their AMT participants were slightly slower than those of the laboratory participants (e.g., Proctor & Lu, 1999, Experiment 1). The RTs of the present participants fell between the AMT and laboratory results. In addition, the RTs of the present participants were slightly slower than those obtained from Japanese students in a laboratory setting (e.g., Watanabe & Yoshizaki, 2016). Furthermore, the student participants responded relatively more quickly and less accurately than the CW participants.

**Experiment 5: Inhibition of Return**

Attention to certain objects in peripheral locations can be facilitated by the preceding stimuli near a location. However, once attention is removed, the processing of the object displayed later at the same location will be hindered. This inhibition of the processing of the previously cued target is referred to as the “inhibition of return” (IOR; for example, Klein, 2000). To examine this effect, researchers typically administer a visual-cueing task. In a typical visual-cueing task, the participants are asked to detect a target stimulus that is located either on the left or right side of the screen. Prior to displaying the target, the cueing stimulus appears at either near the target location (valid cue) or in a different location (invalid cue). Previous investigations using the visual-cueing paradigm have demonstrated that if the delay between displaying the cue and target (cue-target onset asynchrony; CTOA) is short (e.g., ≤300 ms), valid cues result in quicker detection of the target compared with invalid cues (Klein, 2000). However, when the CTOA is longer, the valid cue hinders faster target detection. In addition to a simple detection task, IOR also occurs in a location discrimination task (e.g., Y. Tanaka & Shimojo, 1996).
Participants and procedure. One hundred students (48 female, 91% right-handed, mean age = 19.8 years) and 97 workers (51 female, 95% right-handed, mean age = 38.5 years) participated in the experiment. In each trial, the participants were presented with three placeholder squares and fixation crosses in the center square for 1,000 ms, followed by a 100-ms cue that appeared in the left or right square. After the cue was removed, one of four CTOAs was inserted, followed by a target (black ‘‘X’’) that was displayed either in the left or right placeholder. The participants were asked to detect ‘‘X’’ and indicate its location by pressing one of two keys as quickly and accurately as possible. When the participants pressed the key, the target and placeholders were removed immediately, followed by an accuracy feedback display for 500 ms.

Figure 5. Mean RTs and error rates as a function of cue validity and cue-target onset asynchrony with standard error bars. Note. RTs = reaction times; ERs = error rates; CTOA = cue-target onset asynchrony.

Method

Stimuli and design. A similar procedure and stimulus set as that in Crump et al. (2013) was applied to this task, except that participants were required not only to detect a target but also to discern its location. Three 150-px white-filled squares with black borders appeared as placeholders on white backgrounds. Each placeholder was separated by 170 px. The target was a black ‘‘X’’ that was approximately 25 px in width × 30 px in height and was displayed in the center of either the right or left placeholder. The participants were asked to identify the target location and press the F key if the target appeared in the left placeholder square or the J key if the target appeared in the right square. The cues were 100-px gray squares that appeared in either the right or left placeholder squares. The fixation cross always appeared in the center placeholder. The experimental design involves a 2 (cue validity; valid or invalid) × 4 (CTOA; 200, 500, 900, 1,100 ms) factorial design.

A total of 96 trials consisting of 48 valid-cue and 48 invalid-cue trials (12 trials for each CTOA) were performed by each participant. Valid cues were displayed in the placeholder in which the target appeared, and invalid cues appeared in the opposite direction. The location of the target was counterbalanced. For each trial, the target location, cue validity, and CTOA were assigned randomly.

Participants and procedure. One hundred students (48 female, 91% right-handed, mean age = 19.8 years) and 97 workers (51 female, 95% right-handed, mean age = 38.5 years) participated in the experiment. In each trial, the participants were presented with three placeholder squares and fixation crosses in the center square for 1,000 ms, followed by a 100-ms cue that appeared in the left or right square. After the cue was removed, one of four CTOAs was inserted, followed by a target (black ‘‘X’’) that was displayed either in the left or right placeholder. The participants were asked to detect ‘‘X’’ and indicate its location by pressing one of two keys as quickly and accurately as possible. When the participants pressed the key, the target and placeholders were removed immediately, followed by an accuracy feedback display for 500 ms.

Results and Discussion

No participants were excluded due to low accuracy, but the outlier analysis excluded 0.9% of the observations (UNIV = 0.7 vs. CW = 1.0 trials per participant).

The mean RTs and error rates for each participant were submitted to separate 2 (sample) × 2 (cue validity) × 4 (CTOA) mixed-design ANOVAs. When the assumption of sphericity was violated, the Greenhouse-Geisser corrections were applied to the following analyses. We reported the corrected $p$ values and uncorrected degrees of freedom in these cases. The results showed a significant main effect of the sample on the mean RTs, $M_{\text{UNIV}} = 402$ ms ($SD = 69.6$), $M_{\text{CW}} = 461$ ms ($SD = 105.5$), $F(1, 195) = 24.9$, $p < .001$, $\eta^2_p = .11$. Although these results indicated that the students responded more quickly than the CW participants, we failed to find the main effect of the sample on the error rates or any interaction effects, including that of the sample on both the mean RTs and error rates ($F$s < 1.93, $ps > .13$), except for a marginal Sample × Cue validity interaction, $F(1, 195) = 3.7$, $p = .054$, $\eta^2_p = .02$. Therefore, we will not discuss the sample difference hereafter.

Figure 5 exhibits the mean RTs and error rates as a function of cue validity and CTOA. The analysis of the mean RTs showed a significant main effect of CTOA, $F(3, 585) = 191.4$, $p < .001$, $\eta^2_p = .50$. The post hoc comparisons revealed that all combinations of CTOA pairs, except for 900−1,100 ms, were significantly different from one another (all $ps < .001$). The main effect of the cue validity was also significant, $M_{\text{valid}} = 434$ ms ($SD = 96.5$) vs. $M_{\text{invalid}} = 429$ ms ($SD = 96.1$), $F(1, 195) = 5.5$, $p = .02$, $\eta^2_p = .03$. The Validity × CTOA interaction was also significant, $F(3, 585) = 10.2$, $p < .001$, $\eta^2_p = .05$. The analyses of the simple main effect indicated that the participants detected valid-cue targets more quickly in short CTOAs (200 ms), $F(1, 195) = 10.1$, $p = .002$, $\eta^2_p = .05$; however, they required more time in longer CTOAs (500, 900 or 1,100 ms), $F(3, 195) = 5.2$, $8.9$, $12.7$, $p = .024$, <.003, <.001, $\eta^2_p = .03$, .04, .06, respectively.

For the error rate, the main effect of CTOA was significant, $F(3, 585) = 8.2$, $p < .001$, $\eta^2_p = .04$. The post hoc comparisons revealed that the error rates of the 200 ms CTOA were higher than those of the 1,100 and 900 ms
CTOA, \( ps = .025 < .001 \). The main effect of the cue validity was also significant, \( M_{\text{valid}} = .006 (SD = 0.025) \) versus \( M_{\text{invalid}} = .010 (SD = 0.033) \), \( F(1, 195) = 8.9, p = .003, \eta^2_p = .04 \). The Validity \times CTOA interaction was not significant, \( F(3, 585) = 1.8, p = .149 \).

To summarize, the present results successfully replicated the positive cueing effects in the short CTOA (<300 ms) and negative cueing effects in the longer CTOA (≥ 400 ms). The mean RTs ranged between the detection and discrimination task paradigms in the laboratory experiment (e.g., Lupián, Milán, Tornay, Madrid, & Tudela, 1997).

**General Discussion**

**Characteristics of CW Workers as Participants of Cognitive Behavior Studies**

Recent studies with AMT as a participant pool have suggested that Internet-based behavioral experiments work well, even in cognitive tasks that require precise millisecond control for stimulus presentation and response collection within multiple trials (Crump et al., 2013). The present study attempted to extend these findings to a Japanese crowdsourcing (i.e., non-Caucasian) sample. The results appeared to be in line with the previous findings that were obtained in the laboratory experiments and from the AMT sample. The present results suggested that the Japanese crowdsourcing service is another viable participant pool for cognitive behavior studies.

The present results also demonstrated a rather strong and systematic sample difference in which students were faster but more inaccurate in responding to conflicting stimuli than workers. The supplemental analysis (see Table S2) also indicated that the number of incorrect trials was higher in students, whereas the number of outliers was higher in workers. These results may suggest a trade-off between the accuracy and speed of the responses. One possible explanation is that students were less attentive and impulsive compared with workers. Recently, Hauser and Schwartz (2016) revealed that AMT workers showed significantly better performance in attentional check questions than participants from the student subject pool. The reasons why students paid less attention than workers have not been fully investigated yet; however, certain differences between the two samples, such as monetary compensation and age, may affect attentiveness in participating in the psychological experiments.

One of the major differences between the two samples is that the CW workers were significantly older than the students. Therefore, it is reasonable to assume that the participants became more attentive and cautious as they aged; as a result, their responses became slower but more accurate compared with those of the younger participants. In other words, the differences in RTs and error rates may reflect the difference in age but not the other qualitative differences between students and workers. We then conducted a series of supplemental analyses of covariance (ANCOVAs) and included age as a covariate. These analyses showed that the two samples were not different in their response accuracy and speed after controlling for age (detailed results are shown in the supplemental materials). The results suggested that the differences in accuracy may not have resulted from qualitative differences between the two samples and that the participants could become more attentive and cautious with age.

It is also possible that the speed of the applicable Internet connection affects the participants’ performance. For example, the unfinished rates of the task were higher for the CW worker than for students (10.3% vs. 1.3%; see Table S1 in the supplemental materials). The experimental task adopted here collected data on the speed of the Internet connection at the beginning of the survey. When the connection speed was too slow, the participants were asked not to take the survey and were directed to the end of the survey. Although many student participants accessed the survey page from their universities, the worker participants seemed to access the page from their homes. Generally, universities offer high-speed and stable Internet connections to their communities; therefore, students were less likely to be excluded because of slow connections (see Table S5 in the supplemental materials). High-speed Internet connections may also result in faster RTs, as is particularly found in the student sample. To see this, we calculated the correlation coefficients between the individual’s connection speed and the performance of the tasks (for additional details, see Table S6 in the supplemental materials). The results indicated that a slower Internet connection does not result in higher error rates but instead in slower RTs.

In addition, the unfinished rate of the task may be affected by the compensation to the participants. In the present study, the students received extra course credit while workers were paid for their contributions. A recent study suggested that the primary reason for performing tasks has shifted from intrinsic motivations, such as enjoyment and intellectual curiosity, to extrinsic motivation, that is, financial rewards (Litman, Robinson, & Rosenzweig, 2015; see also Gleibs, 2016). Some workers may then abandon the task when they feel that participating in the survey did not pay off.

Future research should investigate whether and how qualitative differences between traditional and crowdsourcing subject pools, such as age distribution, affect the underlying mental processes when completing the cognitive experimental tasks. It is also important to explore how environmental factors (e.g., connection speed, amount of monetary compensation) enhance and/or impair the performance of participants in the online cognitive experiments.

**Differences Between Caucasian and Non-Caucasian Samples**

The results also demonstrated that the RTs of the present participants were systematically faster than those of the AMT workers (e.g., Barnhoorn et al., 2015; Crump et al., 2013; see...
Table S4 in the supplemental materials). We cannot offer a satisfactory explanation for this result; however, it is possible that language proficiency may play a role. As almost all of the present participants were native Japanese speakers, it is highly unlikely that they had any difficulties performing the designated tasks. However, as Goodman et al. (2013) noted, “English as a Second Language” speakers were likely to fail attentional check question written in English. Therefore, it may be possible that some ESL speakers in the previous study demonstrated a slower response speed due to language proficiency. As demographic information was not collected in the previous study, the effect of language proficiency on performance in the simple cognitive behavioral experiments remains open.

Another possibility is that faster RTs of the present participants compared with the AMT worker may be due to differences in the speed of the Internet connection for the two countries. According to the report (Akamai Technologies, 2015), the average connection speeds were faster in Japan (Japan = 15.0 Mbps, the United States = 12.6 Mbps). As mentioned previously, our supplemental analyses revealed that the server communication delay was associated with longer RTs; therefore, the faster RTs found with the Japanese participants may be due to the relatively fast Internet connections available to them.

However, another possibility is that cultural differences in cognition, particularly in the sensitivity to context, could explain the differences. Recent cross-cultural psychology often differentiates between Eastern “holistic” and Western “analytic” cognitive styles (Nisbett, Choi, Peng, & Norenzayan, 2001) and found that people from both cultures are different not only in their higher cognitive processes but also in relatively lower processes, such as judgment of line length (Kitayama, Duffy, Kawamura, & Larsen, 2003) and visual illusion (Doherty, Tsuji, & Phillips, 2008). These studies have suggested that East Asian participants are likely to perceive a focal object within its contexts; hence, they adopt context-dependent cognitive styles. On the contrary, Western participants tend to dissociate the focal object from environment and hence are context independent. Whether the cultural difference in context sensitivity plays a major role in processing simple but conflicting stimuli remains unclear; however, a few previous studies mentioned considerable cultural differences in processing such stimuli. For example, Kitayama and Ishii (2002) adopted a modified Stroop paradigm to investigate the interference effect of the vocal tone of speech on judgments on the meaning of an emotional word. They found that the interference was stronger for native Japanese speakers; on the contrary, native English speakers had greater difficulty evaluating vocal tone when the content of the word conflicted with vocal emotion. The present study did not aim at a cross-cultural comparison itself; however, an online cognitive experiment is a viable approach to cross-cultural studies, including cognitive behavior experiments.

Limitations and Suggestions

The present study suggested that data collection using a non-Caucasian crowdsourcing sample remains a viable approach in psychology and related research fields that adopt a cognitive behavior experimental paradigm. Despite the many benefits of crowdsourcing, researchers should also consider the following points. First, the participants must be paid fairly, as suggested by previous studies (e.g., Crump et al., 2013; Mason & Suri, 2012). AMT workers are sometimes paid very little (e.g., $.10) for completing simple HITs; however, such rates are lower than those applied to laboratory research. In this study, compensation (i.e., 50 JPY for a 5-min experiment) was decided based on the median value of other academic survey studies found in CrowdWorks. This rate was slightly less than the lowest wage of paid workers in the local city (764 JPY per hour). Nonetheless, it was also shown that the amount of compensation did not impair the data quality (e.g., Buhrmester, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014). The amount of compensation that is considered ethical to complete a study may depend on the task complexity, estimated time of completion, the minimum wage of other paid workers, and so forth; however, it may be time for the research community to issue guidelines for ethically valid compensation for participation in studies.

On the limitation side of the present study, it is important to note an issue of language used by CrowdWorks participants. At this moment, CrowdWorks provides only a user interface written in Japanese; hence, the researchers and participants had to be literate in Japanese. A recent survey showed that more than 70% of Japanese individuals estimated their English comprehension skills as low, for example, “only know simple words” or “cannot read at all” (Koiso, 2009). Therefore, if researchers wish to collect data via Japanese crowdsourcing services, they must prepare all materials in Japanese. It is apparent that an obstacle to conducting cognitive behavioral studies is the use of Japanese participants by researchers who are not literate in Japanese; however, this represents a good opportunity to facilitate cooperative research projects by scientists with different cultural backgrounds. Of course, the same is true of other crowdsourcing services in which most potential participants are not literate in English.

In addition, the previous study cited the limitation in browser-based display technologies, mainly with regard to precise millisecond control for stimulus presentation in quite short periods (less than 80 ms; Crump et al., 2013). The present study did not adopt tasks in which quite fast and precise control was required (e.g., attentional blink experiment); however, it suggests that the slow Internet connection may result in slower RTs. The idea that longer RTs are partly due to delays in communicating with servers may explain why the present student sample is generally slower in responding than those from previous laboratory studies with Japanese students. In the laboratory settings, one can hardly expect any delays in communication unless the experiment is conducted with a
low-end PC on a highly loaded condition. Therefore, the laboratory participants are likely to record faster RTs compared to online studies. Future studies should investigate whether and how the speeds of the Internet connection and/or computational load on PC affect performance in greater detail.

Recently, Majima et al. (2017) suggested that Japanese crowdsourcing workers are eligible for behavioral research that collects data with online questionnaires. The present study provides further findings for this line of research. Our results suggest that data collection using a non-Caucasian crowdsourcing pool remains a promising approach for experimental cognitive research. Generally, our tests indicated that the data were fairly reliable and were consistent with previous laboratory studies. Although certain ethical and technical issues remain open, data collection using crowdsourcing may be essential for promoting empirical behavioral research.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Statement
The experiments reported here were approved and are in compliance with the guidelines of the Hokusei Gakuen University Ethics Committee.

Funding
The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This research was financially supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (15K04033), and the Special Group Research Grant from Hokusei Gakuen University.

Notes
1. Majima, Nishiyama, Nishihara, and Hata (2017) and the present study adopted the same recruitment method; hence, there would be an overlap of participants between two studies. However, we cannot estimate the degree of overlap because we did not track any identity information to keep anonymity of participants.
2. The QRTEngine is an open-source library designed to be embedded in the existing survey suite Qualtrics. This engine was written in Javascript and developed for researchers without substantial programming skills. It does not require the installation of additional software and browser plugins. It is important to note that the QRTEngine team has recently decided to retire the engine because of compatibility issues between the engine and Qualtrics environments, and recommended to finish using QRTEngine (Steenbergen, 2016). The engine still seems to work under the old Qualtrics environment in October, 2016; however, it is likely that the old environment is no longer available in the near future. Of course, there are other tools, such as jsPsych (de Leeuw, 2015); hence, it is still possible to conduct online cognitive behavior studies with these tools.
3. Barnhoorn, Haasnoot, Bocanegra, and van Steenbergen (2015) provided an example task on their website (http://www.qrtengine.com/downloads/download-example-tasks/).

The present study utilized the Stroop task example and translated the instructions and color words into Japanese. The tasks administered in the present study can be obtained from the Internet (see the supplement material).

Supplemental Material
The supplemental material is available at http://journals.sagepub.com/doi/suppl/10.1177/2158244017698731

References
Akamai Technologies. (2015). Akamai’s [State of the Internet] Q3 2015 Report. Retrieved from https://www.akamai.com/us/en-multimedia/documents/report/q3-2015-ott-connectivity-final.pdf
Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. Behavior Research Methods, 47, 918-929. doi:10.3758/s13428-014-0530-7
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. doi:10.3758/s13428-011-0081-0
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
Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. Annual Review of Clinical Psychology, 12, 53-81. doi:10.1146/annurev-clinpsy-021815-093623
Clifford, S., Jewell, R. M., & Waggner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? Research & Politics, 2(4). doi:10.1177/2053168015622072
Crump, M. J., 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
de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. Behavior Research Methods, 47, 1-12. doi:10.3758/s13428-014-0458-y
Doherty, M. J., Tsuji, H., & Phillips, W. A. (2008). The context sensitivity of visual size perception varies across cultures. Perception, 37, 1426-1433. doi:10.1068/p5946
Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception & Psychophysics, 16, 143-149. doi:10.3758/bf03203267
Fukuhara, M., Miki, M., Yokouchi, H., & Hiroyasu, T. (2013). Stroop test ji ni okeru nouketsuryuu henka to kadai seiseki no kankeisei no kento [Discussion of the relation between the cerebral blood flow and reaction time during Stroop test]. The Science and Engineering Review of Doshisha University, 53, 185-190. Retrieved from https://doors.doshisha.ac.jp/duar/repository/ir/15636/023053040003.pdf
Gleibs, I. H. (2016). Are all “research fields” equal? Rethinking practice for the use of data from crowdsourcing market places. Behavior Research Methods. Advance online publication. doi:10.3758/s13428-016-0789-y
Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. Journal of Behavioral Decision Making, 26, 213-224. doi:10.1002/bdm.1753
Harada, E., Asano, A., Suto, S., & Hasher, L. (2010). Kadai kireikae ga motarasu sokushin kouka to karei: Suuji sentaku kadai ni yoru kentou [Facilitative effects of task switching and cognitive aging item RT analysis in the Numerical selection task]. Proceedings of the Japanese Society for Cognitive Psychology, 2010, 46-46. doi:10.14875/cogpsy.2010.0.46.0

Hauser, D. J., & Schwartz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behavior Research Methods, 48, 400-407. doi:10.3758/s13428-015-0578-z

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. Psychological Bulletin, 136, 849-874. doi:10.1037/a0019842

Kitayama, S., Duffy, S., Kawamura, T., & Larsen, J. T. (2003). Perceiving an object and its context in different cultures: A cultural look at new look. Psychological Science, 14, 201-206. doi:10.1111/1467-9220.02432

Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. Cognition & Emotion, 16, 29-59. doi:10.1080/0269993004300121

Klein, R. M. (2000). Inhibition of return. Trends in Cognitive Sciences, 4, 138-147. doi:10.1016/S1364-6613(00)01452-2

Koiro, K. (2009). Nihonjin Eigo Shiyousha no Tokuchou to Eigo Nouryoku. JGSS-2002 to JGSS-2006 no data kara [The characteristics of English users and English proficiency of Japanese people: From the data of JGSS-2002 and JGSS-2006]. JGSS Research Series, 6, 123-137. Retrieved from http://jgss.daishodai.ac.jp/research/monographs/jgssm9/jgssm9_08.pdf

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. Behavior Research Methods, 47, 519-528. doi:10.3758/s13428-014-0483-x

Lupiñanez, J., Milán, E. G., Tornay, F. J., Madrid, E., & Tudela, P. (1997). Does IOR occur in discrimination tasks? Yes, it does, but later. Perception & Psychophysics, 59, 1241-1254. doi:10.3758/bf03214211

Lutz, J. (2016). The validity of crowdsourcing data in studying anger and aggressive behavior. Social Psychology, 47, 38-51. doi:10.1027/1869-9335/a000256

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. Psychological Bulletin, 109, 163-203. doi:10.1037/0033-2909.109.2.163

Majima, Y., Nishiyama, K., Nishihara, A., & Hata, R. (2017). Conducting online behavioral research using crowdsourcing services in Japan. Frontiers in Psychology. doi: 10.3389/ fpsyg.2017.00378

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s mechanical Turk. Behavior Research Methods, 44, 1-23. doi:10.3758/s13428-011-0124-6

Matsuda, Y. (2012). Fukki yokusei to chuu ni kanse ni yoru sokushin kouka [Inhibition of return and facilitation effect by attentional momentum]. Journal of Human Environmental Studies, 10, 109-114. doi:10.4189/shes.10.109

Monsell, S. (2003). Task switching. Trends in Cognitive Sciences, 7, 134-140. doi:10.1016/S1364-6613(03)00028-7

Nisbett, R. E., Choi, I., Peng, K., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. Psychological Review, 108, 291-310. doi:10.1037/0033-295X.108.2.291

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical Turk as a participant pool. Current Directions in Psychological Science, 23, 184-188. doi:10.1177/0963721414531598

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

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2015). Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online research. NA-Advances in Consumer Research, 43, 18-22.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.), Attention and performance X (pp. 531-556). Hillsdale, NJ: Lawrence Erlbaum.

Proctor, R. W., & Lu, C.-H. (1999). Processing irrelevant location information: Practice and transfer effects in choice-reaction tasks. Memory & Cognition, 27, 63-77. doi:10.3758/bf03201214

Selst, M. V., & Jolicour, P. (1994). A solution to the effect of sample size on outlier elimination. Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 47, 631-650. doi:10.1080/1460244040841131

Simon, J. R., Sly, P. E., & Vilapakkam, S. (1981). Effect of compatibility of S-R mapping on reactions toward the stimulus source. Acta Psychologica, 47, 63-81. doi:10.1016/0001-6918(81)90039-1

Steinberg, H. (2016, June 16). Retiring the ORTEngine [Web log post]. Retrieved from http://www.qrtengine.com/retiring-the-qrtengine/

Stoop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643-662. doi:10.1037/b0054651

Tanaka, H., Masaki, H., Takasawa, N., & Yamazaki, K. (2002). Error shori no chuu shigen haibun to error kanren insei denni [The relationship between the error-related negativity (ERN) and mental resource allocation in the error-related processes]. Japanese Journal of Physiological Psychology and Physiopsychology, 20, 29-37. doi:10.5674/jjppp1983.20.29

Tanaka, Y., & Shimojo, S. (1996). Location vs feature: Reaction cost: An investigation using lateralized readiness potentials. Japanese Journal of Psychology, 69, 849-855. doi:10.2391/jjpsy.56.626

Umebayashi, K., & Okita, T. (2011). Kirikae cost no hitaihousei: Hensokusuei junci denni ni yoru kentou [Asymmetric switch cost: An investigation using lateralized readiness potentials]. The Japanese Journal of Psychology, 82, 16-23. doi:10.4992/jjpsy.82.16

Watanabe, Y., & Yoshizaki, K. (2016). Jikkou seigyo no fuka ga Simon kouka ni shiysaya ni oyobosu eikyou [Effect of task load on lateral difference in the Simon task]. Bulletin of Aichi Shukutoku University Faculty of Psychology, 6, 55-62. Retrieved from http://aska-r.aasa.ac.jp/dspace/bitstream/16038/5767/1/0038006201603055062.pdf

Yamazaki, A. (1985). Stoop kannshou ni okeru shorikaietei [Interference in the Stroop color-naming task]. The Japanese Journal of Psychology, 56, 185-191. doi:10.4992/jjpsy.56.185

Author Biography

Yoshimasa Majima is an associate professor of cognitive psychology at the Hokusei Gakuen University. His research focuses on human reasoning and judgment, particularly the process of acquiring everyday beliefs, judgement based on those beliefs and the cultural differences in thinking and cognition.